Thèse Professionnelle
Mastère Spécialisé Big Data

Télécom ParisTech

Contribution au développement d'outils d'apprentissage automatique et d'aide à la décision dans le processus de gestion et d'optimisation des levées de capitaux à destination des sociétés privées

---

Kaëlig Castor

Octobre 2020 - Mars 2021

Praexo,
39 Rue d'Aboukir,
75002 Paris

Remerciements

Cette étude a été effectuée dans le cadre du Mastère Spécialisé Big Data de Télécom ParisTech.

Ce programme m'a énormément apporté. Par conséquent, je remercie toutes les personnes, aussi bien les professeurs que le personnel administratif, qui font un travail de grande qualité et qui ont permis à cette formation de se dérouler dans de bonnes conditions, malgré une situation exceptionnellement difficile cette année, avec le déménagement de l'école sur le plateau de Saclay, les grèves de transport, et la crise sanitaire.

Je remercie Guillaume Moinet, CEO de Praexo pour sa confiance et son accueil au sein de la société, à l'origine de ce travail. Je tiens à remercier également Laurent Lepinay, Head of Development et Jean-David Lavaud, CTO pour leur aide et leur accompagnement.

Je remercie Alexis Perrier, Chief Data Officier, pour son expérience et ses précieux conseils. Je remercie également Fabien Vavrand, data scientist avec qui j'ai collaboré pendant quelques semaines. Je remercie vivement Zouhair Khatouri, Axel Yana, et Yassine Boujerfaoui, stagiaires de l'équipe de data science, avec qui j'ai pu travailler dans les meilleures conditions. Enfin, je remercie chaleureusement toute l'équipe de Praexo pour son accueil sympathique et sa bonne humeur au quotidien.

Résumé

Dans ce rapport sont présentées des contributions à la réalisation d'outils de traitement automatisé des données dans le processus de levées de fonds d'entreprises. L'objectif est de fournir des analyses fiables d'aide à la décision dans le cadre de la relation entre les sociétés et leurs actionnaires et investisseurs. Deux projets ont été abordés. Le premier consiste en un système de recommandation d'investisseurs en fonction de toutes les caractéristiques de transaction telles que le montant, les secteurs d'activité ou géographique, mais aussi des caractéristiques du profil des investisseurs. Trois méthodes distinctes ont été testées pour effectuer le système de recommandation. Plus succinctement, le second projet porte sur un système de synthèse et de catégorisation automatique des commentaires des analystes.

Table des matières

Table des figures

Chapitre 1

Introduction

Dans ce chapitre, le contexte des levées de capitaux est présenté en introduisant quelques notions financières. Pour favoriser la compréhension du projet, les objectifs business sont également exposés avec une présentation de la société Praexo.

1 Les transactions de capitaux privés

Pour financer ses investissements et ses dépenses d'exploitation, une entreprise peut utiliser soit ses capitaux propres, soit des emprunts. Les capitaux propres sont des moyens de financement que l'entreprise n'a pas à rembourser mais qui peuvent cependant être rémunérés via le versement de dividendes qui proviennent soit de l'incorporation dans le capital d'une partie des bénéfices, soit de la levée de capitaux nouveaux. Ces derniers peuvent être trouvés auprès d'investisseurs individuels, personnes physiques ou morales qui acquièrent directement une partie du capital (*private equity*). Ils peuvent aussi être levés auprès du public, via l'introduction en Bourse et l'émission d'actions. Le capital-risque a le potentiel de générer des rendements élevés, mais un investissement dans une entreprise en phase de démarrage est intrinsèquement risqué. Aussi les besoins de transparence dans la communication entre les investisseurs et les entreprises non financières (*corporate*) sont essentiels dans le fonctionnement d'une économie de marché.

2 La technologie au service des émetteurs

Les émetteurs sont les agents, publics ou privés, en quête de ressources destinées à financer leurs investissements : les entreprises, les Etats, les collectivités locales. Le rôle joué par les émetteurs dans l'évolution des marchés est souvent occulté, par rapport aux acteurs phares que sont les banques d'investissement et les investisseurs institutionnels. Pourtant, la déréglementation a permis aux émetteurs de diversifier leurs sources de financement, et de ce fait, de mettre celles-ci en concurrence : c'est bien aussi sous la pression de la demande que l'innovation financière progresse.

Il a été observé, depuis quelques mois, de nouvelles techniques qui transforment la manière dont les entreprises lèvent des capitaux, parmi lesquelles on peut citer :

- l'offre de jetons de titres (ou *STO i.e. Security Token Offerings*) qui répond à un cadre législatif et réglementaire auprès de la SEC (*Securities and Exchange Commission*), l'organisme fédéral américain de réglementation et de contrôle des marchés financiers, ou bien d'autres régulateurs stricts.
- la cotation directe (*direct listing*), moins coûteuse en frais d'introduction par rapport à une introduction classique (ou *IPO i.e. Initial Public Offering*), qui est utilisée quand des

actionnaires historiques souhaitent céder tout ou partie de leurs titres sur le marché sans augmentation de capital.

- les SPAC (*Special Purpose Acquisition Company*) qui sont des entreprises sans activité opérationnelle dont le but est de lever des fonds en entrant sur une place boursière en vue d'une acquisition ou d'une fusion future dans un secteur particulier et avant une échéance déterminée.

Ces types de transactions partagent toutes en commun une chose : une plus forte interaction entre les émetteurs et leurs investisseurs. Le phénomène s'est accéléré dans les récents mois, pendant la pandémie de COVID-19, avec le développement et l'accélération de tournées virtuelles de présentation (*online financial roadshows*) dans les marchés de capitaux [Dhillon et al., 2020].

3 Préparer les entreprises à leur prochain financement

Lorsqu'une société se développe et passe par différentes étapes de financement, le groupe d'investisseurs ainsi que leurs attentes tendent à évoluer avec un besoin toujours plus important de complément d'information et de communication fréquente. En particulier, il y a un intérêt grandissant des investisseurs transverses (*"crossover investors"*) pour participer aux levées pré-publiques. Pourtant ces investisseurs ressentent parfois que les entreprises privées peinent à développer les compétences essentielles et comprendre comment développer leur récit et communiquer efficacement avec les différents profils d'investisseurs notamment ceux qui diffèrent des investisseurs de capital risque.

Réaliser des levées de capitaux dans d'excellentes conditions est un objectif clé dans le succès et la croissance d'une entreprise. La solution de PRAEXO est conçue de sorte que les compagnies et leur management aient une opportunité unique de mieux comprendre les attentes des investisseurs, tout en permettant à ceux-ci de gagner en visibilité pour leur investissement.

4 Présentation de la société Praexo

PRAEXO est une société créée en 2019 par Guillaume Moinet qui a fait toute sa carrière en banque d'investissement au service des sociétés dans le cadre de la réalisation de leurs opérations de levées de capitaux. À la fin décembre 2020, la société employait 10 personnes en CDI, CDD, stage et contrats *freelance*. La société est financée en fonds propres à hauteur de 1.6 millions d'euros apportés par le management et deux investisseurs, ainsi que 0.2 millions d'euros prêtés par la BPI.

PRAEXO propose une solution digitale de type SaaS (*Software as a Service*) destinée à répondre à un besoin de digitalisation et modernisation de l'accompagnement des sociétés dans le cadre de leurs relations investisseurs et de leurs levées de capitaux. L'objectif principal de PRAEXO est de créer une communication directe et transparente entre les sociétés et leurs actionnaires et investisseurs tout en organisant de la collecte, du traitement et de la restitution de données afin de fournir au management des analyses fiables d'aide à la décision. Par ailleurs, les solutions de PRAEXO visent à offrir un meilleur ciblage des investisseurs, et une qualification plus solide du cas d'investissement, dans le but d'optimiser les conditions de réalisation des levées de capitaux. Grâce à une digitalisation des process, le programme d'innovation porté par PRAEXO est destiné à révolutionner la manière dont les interactions avec les investisseurs et les levées de capitaux sont réalisées par les sociétés. Par le biais de recoupement de bases de données reposant sur de l'intelligence artificielle et du *machine learning*, les sociétés peuvent alors mieux identifier les investisseurs à cibler et les thèmes à aborder dans la présentation de leur cas d'investissement. En outre, PRAEXO offre une plateforme digitale qui permet aux managers des sociétés de toujours

être en contact avec l'évolution de leur levée, et de pouvoir ainsi mieux anticiper les attentes des investisseurs grâce à un *dashboard* synthétisant les retours des investisseurs.

Il n'existe pas aujourd'hui de solution digitale complète dans le monde de l'accompagnement des sociétés dans leurs levées de fonds. Les intermédiaires et les banques ont jusqu'à présent très peu utilisé les outils digitaux pour rendre les process plus agiles, plus transparents. Les équipes de Praexo souhaitent mettre en avant une solution de type *"Intelligence Augmentation"* plutôt que *"Artificial Intelligence"* où leur expérience de banquiers conseils couplée aux algorithmes est la seule combinaison efficace dans le cadre des levées de capitaux de type institutionnel. La société se distingue donc de tous les projets de type *crowdfunding* ou plateformes de *tokenisation* qui poussent à la désintermédiation à outrance et qui devrait rencontrer des difficultés pour monter en gamme. Dans la pratique, l'objectif de la société Praexo est d'utiliser les techniques modernes de science des données pour identifier les investisseurs pertinents. Les thématiques et les indicateurs clés de performance (*KPI i.e. Key Performance Indicator*), les plus spécifiques, sont sélectionnés et ordonnés pour préparer un meilleur positionnement de l'entreprise avant qu'il soit présenté aux investisseurs. Tous ces indicateurs sont revus minutieusement par l'équipe sénior de banquiers d'affaires. L'accompagnement des clients est alors optimal puisqu'il combine la double expertise des solutions algorithmiques et des compétences business.

Chez Praexo, il y a une forte conviction du réel mérite des entreprises d'être capable de conduire et gérer une communication directe avec les investisseurs. Le développement de nouvelles technologies permet désormais que de telles interactions puissent être facilement menées directement par le management des sociétés. Grâce aux solutions proposées par Praexo, les entreprises pourront accéder aux investisseurs les plus pertinents à l'échelle globale, et réaliser entièrement leur processus de levée de fonds tout en restant focalisées sur leurs opérations quotidiennes.

5 Identifier les investisseurs les plus pertinents

L'un des objectifs de ce travail est de contribuer au développement d'un outil de recommandation d'investisseurs. L'outil a pour but de réduire le temps passé à des interactions inutiles avec des investisseurs non-pertinents, et de s'assurer que l'information stratégique, et souvent confidentielle, ne soit partagée qu'avec ces investisseurs qui sont probablement le plus à même de participer à une levée de capitaux, et de s'engager dans le processus de croissance de l'entreprise tout au long de ses cycles de développement.

Le système de recommandation d'investisseurs prépare les transactions de marché de capitaux. Avec l'objectif, à terme, de connecter ensemble plusieurs bases de données, et d'appliquer des algorithmes efficaces, ce projet a pour but essentiel de donner aux utilisateurs l'accès à une donnée contextualisée de qualité supérieure dans le but de mieux les préparer et les qualifier à une transaction de marché de capitaux à venir.

Praexo offre un point d'accès unique de recherche à travers des données publiques et propriétaires. Lorsque sont approchés des actionnaires ou de potentiels nouveaux investisseurs d'une entreprise, ce projet intègre en un seul outil de recommandation consistant, une large variété d'informations et de métriques utilisées dans l'industrie financière.

Ces informations fournissent :

- l'historique des transaction de levées de capitaux privés de séries A jusqu'aux introduction en bourse (*Initial Public Offering (IPO)*)
- les listes exhaustives et détaillées des profils d'investisseurs et des investissements passés.

Étant donné le développement ces dernières années d'opportunités d'investissements transverses entre les marchés publics et privés, et l'appétit grandissant des investisseurs pour diversifier leur forme d'investissement, le choix délibéré a été fait chez Praexo de prendre la gamme de types d'investisseurs la plus large incluant les investisseurs institutionnels, les fonds de capital risque (*Venture Capital*) et de capital investissement (*Private Equity*) mais aussi les fonds de capital risque d'entreprises (*corporate VC*) et les sociétés *holding*.

L'objectif n'est pas seulement la création d'une plateforme digitale d'investissement. Toute compagnie cherchant à naviguer dans l'univers de l'investissement réalisera que celui-ci est gouverné par plusieurs dynamiques. Les investisseurs pertinents ne peuvent pas simplement être identifiés sur la base de critères statiques tels que les investissements passés, les types d'investissements, la taille des transaction, ou la situation géographiques des uns et des autres (fig. 1.1). De plus, une compagnie et son management se reposent parfois trop souvent sur leur réseau pour contacter d'éventuels investisseurs. Ils peuvent donc manquer des opportunités d'interagir avec d'autres investisseurs pertinents dont la participation et la contribution peuvent permettre à l'entreprise d'atteindre de nouveaux objectifs dans son cycle de croissance et de développement. Les algorithmes développés par Praexo ont pour but d'être conçus pour identifier les changements et les tendances dans les politiques et les stratégies d'investissement des investisseurs à l'échelle globale.

La digitalisation du processus de levée de capitaux ne doit pas être synonyme de standardisation. Les données issues de l'algorithme de recommandation sont analysées et remises dans leur contexte grâce aux équipes business de Praexo pour aider à optimiser le parcours de l'entreprise dans sa levée. L'objectif est de présenter à chaque investisseur, un positionnement spécifique et adapté à sa manière de fonctionner, en incorporant les *KPI* de l'entreprise les plus appropriés et les plus intéressants pour lui. Ainsi, en étant en accord avec les préoccupations de chaque investisseur, de meilleurs résultats peuvent être obtenus pour l'entreprise dans ses prochains tours de financements, et à terme, de son éventuelle entrée en bourse.

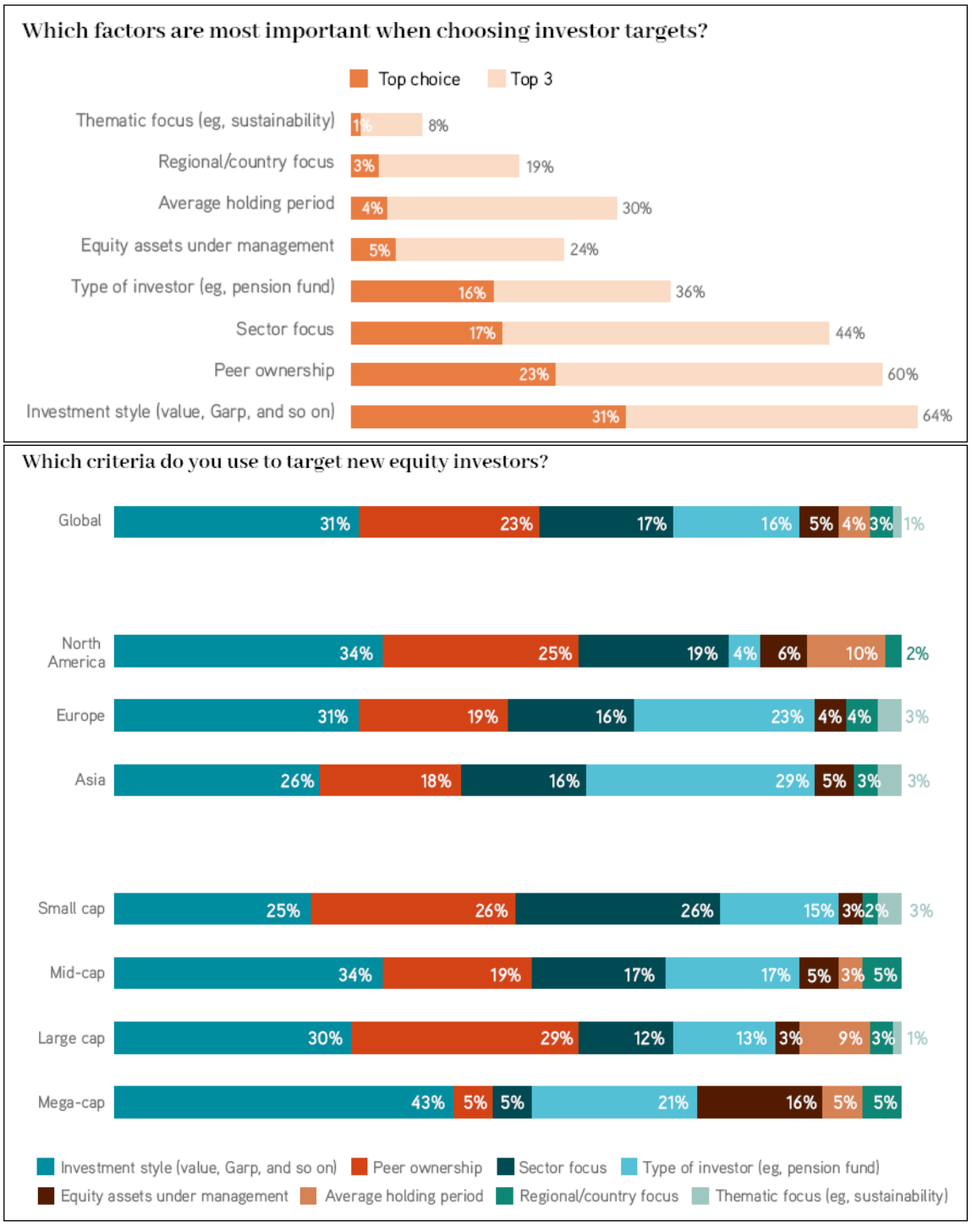

FIGURE 1.1 : Critères de ciblage d'investisseurs [IRmagazine, 2019]

Chapitre 2

Système de recommandation d'investisseurs

En définissant les critères de levée de capitaux d'une entreprise, parmi lesquels se trouvent le secteur d'activité, la taille de la transaction, ou la zone géographique, l'objectif de ce projet est de pouvoir proposer de manière automatique une liste de noms d'investisseurs potentiels, les plus probables, pour cette transaction particulière. Il s'agit donc d'un système de recommandation qui a pour but de sélectionner des investisseurs en filtrant de manière pertinente l'information contenue dans les données des levées de fonds des entreprises. Après une présentation, d'une part, d'un état de l'art des systèmes de recommandation, et d'autre part, de la génération de la base de données, les différentes approches statistiques de *machine learning* effectuées pour la réalisation de cette suggestion d'investisseurs vont être développées.

1 État de l'art

Avec l'abondance de la production de données, de nombreux domaines sont particulièrement touchés par le problème de la surcharge d'information. Les utilisateurs des systèmes d'information sont confrontés à plusieurs problèmes : ils sont submergés par le nombre très important de choix possibles dans l'espace qu'ils explorent. L'exploitation de cette longue liste d'options est complexe et nécessite du temps pour sélectionner celles qui correspondent le mieux aux intérêts de chaque utilisateur. En outre, les utilisateurs ne savent pas nécessairement ce qu'ils devraient voir ou ce qu'ils pourraient apprécier. Leur cheminement n'est alors, en général, pas toujours réfléchi, ou bien ils se limitent à voir les produits les plus populaires dans la plupart des cas. Par conséquent, ils peuvent perdre du temps en sélectionnant des points d'intérêt non pertinents. Inversement, ils peuvent manquer des options qui auraient pu les intéresser. Ainsi, un des domaines de recherche principaux relatifs à la problématique de la surcharge de données est le domaine de la recherche d'information. Le principe général est d'élaborer des méthodes et des algorithmes afin de rechercher des ressources en fonction de requêtes formulées par des utilisateurs. Il n'est cependant pas toujours évident pour un utilisateur de savoir comment exprimer sa demande. De plus, sa requête correspond généralement à une quantité importante de ressources et il est difficile de savoir quels résultats lui présenter en premier, d'autant plus que, d'un utilisateur à un autre, l'ordre de priorité peut changer. Un autre domaine de recherche relatif à cette problématique est donc le domaine des systèmes de recommandation.

Les systèmes de recommandation sont des systèmes de filtrage d'information qui analysent le comportement d'un utilisateur à partir des données disponibles et effectuent ensuite une prédiction des préférences de cet utilisateur, en fonction de ses intérêts, et de la pertinence de l'information. Ils ont commencé à être formalisés à la fin les années 90. Avec l'essor du e-commerce par les plateformes telles Amazon, eBay, ou Shopify, qui impliquent d'effectuer des recommandations de

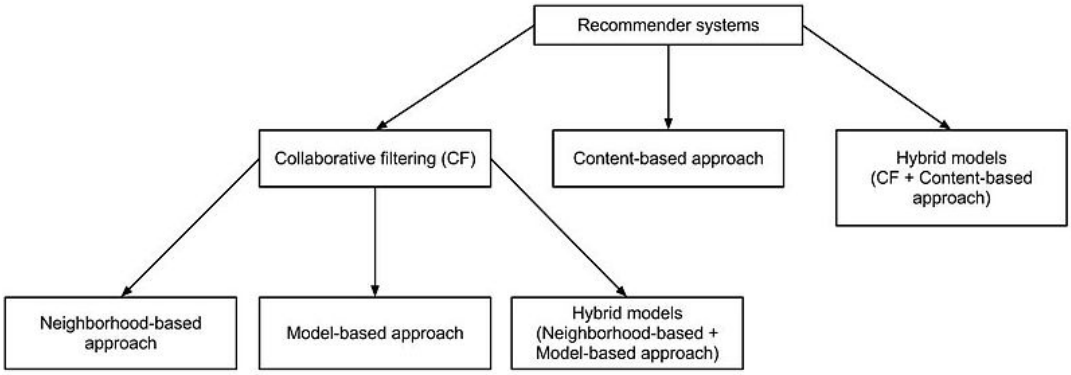

FIGURE 2.1 : Les différents types de Systèmes de Recommandation.

certains produits (*items*) aux utilisateurs (*users*) [Shahbazi et al., 2020], les systèmes de recommandation sont largement utilisés pour aider à naviguer parmi de nombreux choix possibles et surtout pour recommander au mieux des films, des restaurants, des lieux à visiter, ou des articles à acheter. Ces systèmes sont capables de fournir des recommandations adaptées aux préférences et aux besoins des utilisateurs. Ils sont très satisfaisants pour aider les utilisateurs à accéder aux ressources désirées dans un temps limité. Devenus de plus en plus populaires, ils sont aujourd'hui un composant principal de beaucoup d'applications dans différents domaines.

On peut noter qu'un avantage très conséquent des systèmes de recommandation est que l'utilisateur n'a pas obligatoirement besoin de formuler de requête. Sa seule requête peut être implicite et peut se traduire par : "Quelles sont les ressources qui correspondent à mes préférences, mes besoins et mes contraintes ?" [Benouaret, 2017 ; Amer-Yahia and Benouaret, 2020].

Les systèmes de recommandation sont généralement divisés en trois groupes distincts (cf Fig. 2.1) :

- le filtrage collaboratif (*i.e. CF, ou Collaborative Filtering*),
- les recommandations basées sur le contenu (*content based*),
- les méthodes hybrides.

Dans le premier cas, le système basé sur la collaboration utilise les similitudes entre les requêtes et les éléments simultanément pour fournir des recommandations. Il prévoit donc ce qu'un utilisateur particulier aime en fonction de ce que d'autres utilisateurs similaires aiment. Dans le second cas, pour un utilisateur donné, le système basé sur le contenu utilise la similitude entre les éléments pour recommander des éléments similaires à ce que l'utilisateur aime. Il prédit donc un contenu similaire à ce que l'utilisateur a aimé dans le passé. La plupart des entreprises comme Netflix et Hulu utilisent l'approche hybride, qui fournit des recommandations basées sur la combinaison du contenu qu'un utilisateur a aimé dans le passé et de ce que d'autres utilisateurs similaires aiment.

Un système de recommandation est composé de deux étapes :

- L'étape de récupération des données permet la sélection d'un ensemble initial de centaines de candidats parmi tous les candidats possibles. L'objectif principal de ce modèle est d'éliminer efficacement tous les candidats qui n'intéressent pas l'utilisateur. Étant donné que ce modèle d'extraction peut traiter des millions de candidats, il doit être efficace en termes de calcul.
- L'étape de classement prend les sorties du modèle d'extraction et les affine pour sélectionner le meilleur sous-ensemble possible de recommandations. Sa tâche est de restreindre l'en-

semble des éléments susceptibles d'intéresser l'utilisateur à une liste restreinte de candidats potentiels.

Les modèles d'extraction sont souvent composés de deux sous-modèles :

- Un modèle de requête calculant la représentation de requête (normalement un vecteur d'incorporation à dimensionnalité fixe) à l'aide des fonctionnalités de requête.
- Un modèle candidat calculant la représentation candidate (un vecteur de taille égale) à l'aide des caractéristiques candidates.

Les sorties des deux modèles sont ensuite multipliées ensemble pour donner un score d'affinité candidat-requête, des scores plus élevés exprimant une meilleure correspondance entre le candidat et la requête.

Filtrage passif vs filtrage actif

Une fois les données rassemblées, il existe deux méthodes de base pour les filtrer pour faire des prédictions. La méthode la plus élémentaire est le filtrage passif. Pour faire des prédictions, elle utilise simplement des agrégats de données comme la note moyenne d'un élément. La méthode la plus avancée consiste à utiliser le filtrage actif, qui utilise des modèles dans l'historique des utilisateurs pour faire des prédictions. Un exemple de cela serait de trouver des utilisateurs similaires à l'utilisateur actuel et d'utiliser leur historique pour prédire une évaluation.

La distinction entre ces deux méthodes est subtile, mais se résume essentiellement à savoir si les recommandations sont ou non spécifiques à l'utilisateur. Dans le filtrage passif, chaque utilisateur recevra les mêmes prédictions pour un élément particulier. En filtrage actif, le système prend en compte l'historique spécifique de l'utilisateur pour émettre une recommandation. Le filtrage actif est donc ce qu'on appelle un filtrage collaboratif. Bien que le filtrage passif ait des applications très utiles et pratiques, un système de recommandation personnelle ne peut être implémenté qu'en utilisant un filtrage actif.

Filtrage centré sur l'utilisateur et filtrage centré sur l'élément

Tous les systèmes de recommandation doivent décider s'ils tenteront ou non de voir des modèles entre les utilisateurs ou entre les éléments. Un système centré sur l'utilisateur trouvera des similitudes entre les utilisateurs, puis utilisera les préférences des utilisateurs similaires pour prédire les évaluations. Une alternative à cela est un système centré sur les éléments qui tentera de trouver les relations entre les éléments et de faire des prédictions uniquement en fonction des préférences de l'utilisateur et de ces relations.

1.1 Filtrage collaboratif

Les algorithmes de filtrage collaboratif (CF *i.e. collaborative filtering*) ont été très étudiés dans le milieu académique comme dans les entreprises privées. Il y a une hypothèse importante sous-jacente à tout filtrage collaboratif : les utilisateurs qui ont des préférences similaires dans le passé sont susceptibles d'avoir des préférences similaires à l'avenir. C'est cette hypothèse qui permet de prendre l'historique d'un utilisateur et d'extrapoler et de prédire les éléments dont il pourrait bénéficier à l'avenir. Il est évident que cette hypothèse est basée sur le fait qu'il y a des gens qui se ressemblent et d'autres qui sont très différents les uns des autres. Pour les systèmes de recommandation, on examine simplement la similitude de leurs préférences sur un élément particulier.

L'un des principaux avantages de l'approche de filtrage collaboratif est qu'elle ne repose pas sur un contenu analysable par machine et qu'elle est donc capable de recommander avec précision des éléments complexes tels que des films sans nécessiter une "compréhension" de l'élément lui-même.

De nombreux algorithmes, tels que le K-NN ou la corrélation de Pearson, ont été utilisés pour mesurer la similarité des utilisateurs ou la similarité des éléments dans les systèmes de recommandation.

Les algorithmes de filtrage collaboratif peuvent être divisés en deux branches principales :

- basée sur la mémoire,
- basée sur un modèle.

La différence fondamentale est que les algorithmes basés sur la mémoire utilisent toutes les données en permanence pour faire des prédictions, tandis que les algorithmes basés sur des modèles utilisent les données pour apprendre / entraîner un modèle qui peut ensuite être utilisé pour faire des prédictions. Cela signifie que les algorithmes basés sur la mémoire doivent généralement avoir toutes les données en mémoire, alors que les algorithmes basés sur un modèle peuvent faire des prédictions rapides en utilisant moins de données que l'original, une fois que le modèle a été construit. Un exemple bien connu d'approches basées sur la mémoire est l'algorithme basé sur l'utilisateur, tandis que celui des approches basées sur un modèle consiste en un "mappage" de noyau.

Lors de la construction d'un modèle à partir du comportement d'un utilisateur, une distinction est souvent faite entre les formes explicites et implicites de collecte de données.

Collecte de données explicite vs implicite – Afin de faire des recommandations, le système doit collecter des données. Le but ultime de la collecte des données est de se faire une idée des préférences de l'utilisateur, qui peut ensuite être utilisée pour faire des prédictions sur les préférences futures de l'utilisateur [Jannach et al., 2010 ; Zhang et al., 2019 ; Bobadilla et al., 2013].

Il existe deux façons de collecter les données :

- des données de retour utilisateur **explicites** (commentaires, notes, votes, adhésions, souscriptions, préférences, achats, ...),
- des données de retour utilisateur **implicites** (observer, tracer et analyser des comportements sujets à interprétation subtile ou ambivalente tels qu'un historique des actions de l'utilisateur et de leur fréquence incluant ses achats, ses consultations d'articles, ses mouvements de souris, ...).

Les méthodes basées sur les données explicites sont plus répandues et mieux documentées. Elles sont souvent basées sur des notations directes d'utilisateurs pour exprimer leurs préférences concernant des choix spécifiques. La collecte de données explicites est facile à utiliser. On suppose que les évaluations fournies par un utilisateur peuvent être directement interprétées comme les préférences de l'utilisateur, ce qui facilite les extrapolations à partir des données pour prédire les évaluations futures. Cependant, l'inconvénient des données explicites est qu'elles confient la responsabilité de la collecte des données à l'utilisateur, qui peut ne pas vouloir prendre le temps de saisir les évaluations.

D'autre part, les données implicites sont faciles à collecter en grandes quantités sans aucun effort supplémentaire de la part de l'utilisateur. Malheureusement, il est beaucoup plus difficile de travailler avec. L'objectif est de convertir le comportement de l'utilisateur en préférences utilisateur, mais cela nécessite de surmonter un obstacle : comment déduire exactement une préférence basée sur des actions dans un système ? Cela peut être une question à laquelle il est difficile de répondre.

Bien entendu, ces deux méthodes de collecte de données ne s'excluent pas mutuellement. Une combinaison des deux permet d'obtenir les meilleurs résultats : on a les avantages du vote explicite lorsque l'utilisateur choisit d'évaluer des éléments, et, en collectant implicitement des données, on

peut toujours faire des recommandations lorsque l'utilisateur ne note pas les éléments.

Il existe d'autres petites difficultés liées à la collecte de données. La collecte de données ne peut enregistrer que les actions d'un utilisateur et ne sait rien de la personne réelle derrière l'ordinateur. Il est presque impossible de caractériser correctement un utilisateur qui se trouve être en fait deux personnes utilisant le même ordinateur. Les votes explicites peuvent ne pas être une représentation précise des véritables préférences d'un utilisateur si la personne ne se connaît pas assez bien. Certaines études ont montré que les gens peuvent évaluer les articles différemment pour des raisons indiscernables. Enfin, la collecte de données implicites peut impliquer certains problèmes de confidentialité.

La plupart des recherches actuelles améliorent la précision des algorithmes basés sur la mémoire uniquement en améliorant les mesures de similarité. Mais peu de recherches se sont concentrées sur les modèles de score de prédiction qui semblent plus importants pour obtenir de meilleurs résultats de recommandation que les mesures de similarité.

L'algorithme le plus connu en matière de modélisation est la factorisation matricielle. Comparé aux algorithmes basés sur la mémoire, l'algorithme de factorisation matricielle a généralement une précision plus élevée [Delporte, 2014]. Pourtant, la factorisation matricielle peut tomber dans l'optimum local dans le processus d'apprentissage, ce qui conduit à un apprentissage inadéquat.

Comme évoqué précédemment, pour améliorer les performances du système de recommandation, tant du point de vue de la qualité de la recommandation que de la rapidité des calculs, on peut effectuer en amont un clustering des *users* et des *items* pour faire un regroupement au préalable afin de ne pas effectuer la recherche sur l'ensemble des données. Dans la même optique, on peut effectuer de la réduction de dimensions et ne prendre que les *features* les plus significatives.

Les approches de filtrage collaboratif présentent généralement trois types de problèmes :

- Le **démarrage à froid** : il est difficile de traiter un nouveau document ("*new item*") puisque dans l'approche collaborative, les objets à recommander ne sont décrits que par les évaluations fournies par les utilisateurs. De même, pour un nouvel utilisateur, il n'y a pas assez de données pour faire des recommandations précises.
- L'**évolutivité** : dans de nombreux environnements dans lesquels ces systèmes font des recommandations, il existe des millions d'utilisateurs et de produits. Ainsi, une grande puissance de calcul est souvent nécessaire pour calculer les recommandations.
- La **parcimonie** : le nombre d'articles et d'utilisateurs est généralement très important et les notations ne couvrent pas l'ensemble des données, elles ne concernent qu'un sous-ensemble de correspondances éparses.

Le filtrage collaboratif est toujours utilisé dans le cadre des systèmes hybrides.

1.2 Filtrage basé sur le contenu

Les méthodes de filtrage basées sur le contenu sont basées sur une description de l'article et un profil des préférences de l'utilisateur. Ces méthodes sont les mieux adaptées aux situations où il existe des données connues sur un élément (nom, emplacement, description, etc.), mais pas sur l'utilisateur. Les recommandations basées sur le contenu traitent la recommandation comme un problème de classification spécifique à l'utilisateur et apprennent un classifieur pour les goûts et les aversions de l'utilisateur en fonction des caractéristiques d'un élément. Ce type d'approche est intéressant d'un point de vue du temps et des ressources nécessaires pour effectuer les calculs, il donc utile dans un contexte industriel où la rapidité de calcul avec des données de dimensions gigantesques est privilégiée par rapport aux performances du modèle. Cette rapidité de calcul a

un prix et un certain nombre d'informations disponibles ne sont pas mises à profit pour effectuer la recommandation (notamment le comportement des autres utilisateurs).

1.3 Méthodes hybrides

L'intérêt économique lié aux systèmes de recommandation en fait un sujet effervescent et la littérature concernant les nouvelles méthodes est prolifique. On peut citer, par exemple, le récent projet RecBole qui regroupe la plupart des techniques modernes relatives à ce domaine de recherche [Zhao et al., 2020a]. Pour permettre des avancées notables et rapides concernant cette thématique, de nombreux challenges sont organisés chaque année par de grandes entreprises (ACM RecSys2021 Challenge, SIGIR eCom Data Challenge, ACM WSDM2021 Booking.com Challenge, ...). Ainsi, de nombreuses méthodes hybrides ont été développées pour améliorer les systèmes de recommandation [Bobadilla et al., 2020], et il y a un réel besoin de mieux comprendre les différentes forces et faiblesses des algorithmes [Amer-Yahia and Benouaret, 2020]. La figure 2.2 montre les performances de quatre modèles en terme de MAE, MSE, and RMSE. Il s'agit d'une comparaison de la liste des 10 meilleures recommandations issues des algorithmes XGBoost, random forest, support vector regressor, et linear regressor, où l'on voit que XGBoost présente nettement les meilleures performances [Shahbazi et al., 2020 ; Zeinab Shahbazi, 2020].

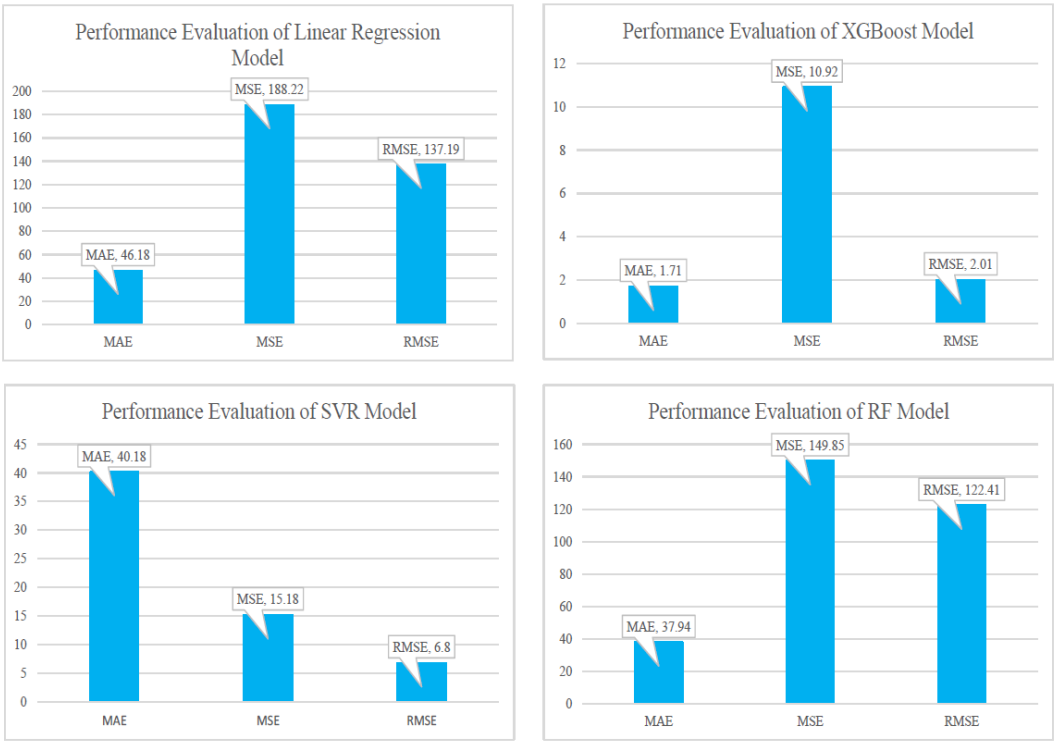

FIGURE 2.2 : Évaluation des performances de quatre différentes techniques d'analyse prédictive.

Les approches des systèmes de recommandation sont particulièrement variées, et peuvent être classées de différentes manières. Toutes ces approches présentent néanmoins des caractéristiques complémentaires. Par conséquent un grand nombre de travaux se sont intéressé à différentes techniques d'hybridation, qui en plus de permettre de profiter des avantages respectifs de ces approches, s'avèrent fournir des recommandations plus précises. Durant ces dernières années, les réseaux de neurones ont connu un large succès pour différentes applications telles que la reconnaissance d'images, de la parole, ou le traitement du langage naturel. Plus récemment, des techniques basées

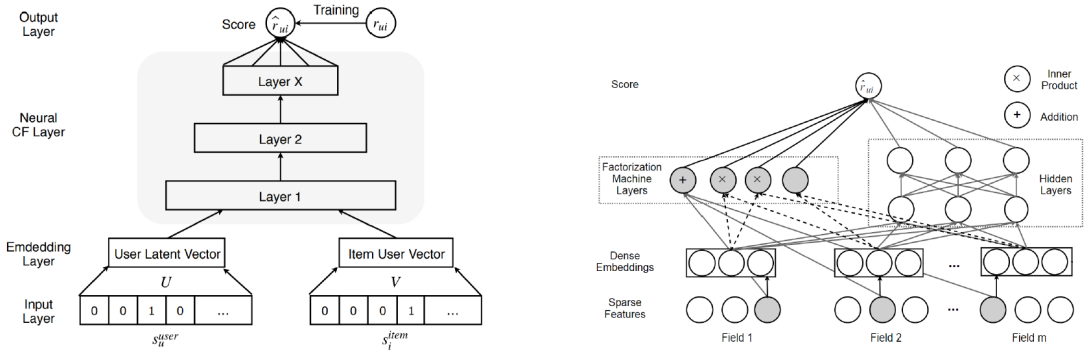

FIGURE 2.3 : Illustration des méthodes : (a) Filtrage collaboratif neuronal (*Neural Collaborative Filtering*) ; (b) Machine de factorisation profonde (*Deep Factorization Machine*) [Zhang et al., 2017 ; Rendle et al., 2020 ; Anelli et al., 2021 ; Xu et al., 2021 ; Cheng et al., 2016]

sur les réseaux de neurones ont vu le jour afin de traiter les problèmes de recommandation [He et al., 2017].

1.4 Réseaux de neurones

L'apprentissage profond présente plusieurs avantages pour réaliser des modèles de recommandation [Zhang et al., 2017]. Il existe de nombreuses méthodes de systèmes de recommandation basés sur les réseaux de neurones profonds comme l'illustre la figure 2.3.

1.4.1 Transformation non linéaire

Contrairement aux modèles linéaires, les réseaux de neurones profonds sont capables de modéliser la non-linéarité dans les données avec des activations non linéaires telles que relu, sigmoïde, tanh, etc. Cette propriété rend possible le fait de pouvoir capturer les intrications et les interactions complexes des données *user-item*. Il est bien établi que les réseaux de neurones sont capables d'approximer n'importe quelle fonction continue en faisant varier les choix d'activation et les combinaisons. Cette propriété permet de gérer des paramètres d'interaction complexes et refléter précisément les préférences de l'utilisateur [Xue et al., 2018]. Les méthodes conventionnelles telles que la factorisation matricielle sont des modèles essentiellement linéaires. Par exemple, la factorisation matricielle modélise l'interaction utilisateur-élément en combinant linéairement l'utilisateur et les facteurs latents de l'élément.

1.4.2 Apprentissage de la représentation

Les réseaux de neurones profonds sont performants pour apprendre l'explication sous-jacente des facteurs et représentations utiles à partir des données d'entrée. En général, une grande quantité d'informations descriptives sur les éléments et les utilisateurs est disponible dans des applications du monde réel. L'utilisation de ces informations fournit un moyen de faire progresser la compréhension des liens entre articles et utilisateurs, favorisant ainsi une meilleure recommandation. L'application des réseaux de neurones profonds à l'apprentissage de la représentation dans la recommandation des modèles permet de traiter la complexité des données multi-modales. Les avantages de l'utilisation de réseaux de neurones profonds pour aider à l'apprentissage de la représentation sont de deux ordres.

- Il réduit les efforts dans la gestion des caractéristiques. L'ingénierie des *features* est un travail intensif, les réseaux de neurones profonds permettent l'apprentissage automatique des fonctionnalités à partir des données brutes de manière non supervisée ou supervisée.
- Il permet aux modèles de recommandation d'inclure des informations de contenu hétérogènes telles que texte, images, audio et même vidéo. Les réseaux d'apprentissage profond ont fait des percées dans le multimédia, traitement des données et ont démontré leur potentiel dans les représentations apprenant à partir de diverses sources.

Espace d'intégration – Le filtrage basé sur le contenu et le filtrage collaboratif mappent chaque élément et chaque requête (ou contexte) à un vecteur d'intégration dans un espace d'intégration commun. En règle générale, l'espace d'inclusion est de faible dimension (beaucoup plus petit que la taille du corpus) et il capture une certaine structure latente de l'élément ou de l'ensemble de requêtes. Des éléments similaires, tels que des vidéos YouTube qui sont généralement regardées par le même utilisateur, se retrouvent rapprochés dans l'espace d'intégration. La notion de "proximité" est définie par une mesure de similarité.

1.5 Mesures de performance des recommandations

Les métriques de recherche d'informations telles que la précision, le rappel ou le Gain cumulatif actualisé (DCG *i.e. Discounted cumulative gain*) sont utiles pour évaluer la qualité d'une méthode de recommandation. La diversité, la nouveauté et la couverture sont également considérées comme des aspects importants de l'évaluation. L'évaluation des performances d'un algorithme de recommandation sur un ensemble de données est extrêmement difficile car il est impossible de prédire avec précision les réactions des utilisateurs réels aux recommandations.

En règle générale, la recherche sur les systèmes de recommandation vise à trouver les algorithmes de recommandation les plus précis. Cependant, il existe un certain nombre de facteurs qui sont également importants dans la mesure de performance.

Diversité – Les utilisateurs ont tendance à être plus satisfaits des recommandations lorsqu'il y a une plus grande diversité intra-liste, par exemple des articles de différents artistes.

Persistance des recommandations – Dans certaines situations, il est plus efficace de réafficher les recommandations, ou de laisser les utilisateurs réévaluer les éléments, que d'afficher de nouveaux éléments. Il y a plusieurs raisons à cela. Les utilisateurs peuvent ignorer des éléments lorsqu'ils sont affichés pour la première fois, par exemple, parce qu'ils n'ont pas eu le temps d'examiner attentivement les recommandations.

Confidentialité – Les systèmes de recommandation doivent généralement traiter des problèmes de confidentialité car les utilisateurs doivent révéler des informations sensibles. La création de profils d'utilisateurs à l'aide du filtrage collaboratif peut être problématique du point de vue de la confidentialité. De nombreux pays européens ont une forte culture de la confidentialité des données, et chaque tentative d'introduire un niveau de profilage des utilisateurs peut entraîner une réponse négative du client. De nombreuses recherches ont été menées sur les problèmes de confidentialité en cours dans cet espace.

Données démographiques des utilisateurs – Les données démographiques des utilisateurs peuvent influencer le degré de satisfaction des utilisateurs vis-à-vis des recommandations.

Robustesse – Lorsque les utilisateurs peuvent participer au système de recommandation, le problème de la fraude doit être résolu.

Sérendipité – La sérendipité se réfère à la notion selon laquelle les recommandations peuvent surprendre l'utilisateur. En recommandant des articles "non-évidents", la sérendipité permet de garantir un effet de suprise dans la liste des suggestions. La sérendipité sert deux objectifs :

- éviter le désintérêt des utilisateurs engendré par des recommandations trop uniformes,
- recommander des éléments "surprenants", nécessaires pour l'apprentissage et l'amélioration des algorithmes.

Confiance et responsabilité – Un système de recommandation n'a que peu de valeur pour un utilisateur s'il ne fait pas confiance au système. Les utilisateurs souhaitent avoir un contrôle sur les recommandations faites et pouvoir indiquer si une recommandation ne leur convient pas. La confiance peut être construite par un système de recommandation en expliquant comment il génère des recommandations et pourquoi il recommande un élément. L'explicabilité est également déterminante pour des questions de responsabilité. En cas d'accident, de discrimination ou toute autre "mauvaise" décision algorithmique, il est essentiel d'auditer le système afin d'identifier la source de la mauvaise décision, corriger l'erreur pour l'avenir, et déterminer les éventuelles responsabilités [Jain and Madhyastha, 2019].

Étiquetage – La satisfaction des utilisateurs à l'égard des recommandations peut être influencée par l'étiquetage des recommandations.

Choix de la métrique pour l'estimation de la performance des prédictions – Dans les systèmes de recommandation, le résultat le plus important est de recevoir une liste ordonnée de recommandations, du meilleur au pire. Dans notre cas, on se soucie pas beaucoup de l'ordre exact de la liste - un ensemble de quelques bonnes recommandations convient. Prenant ce fait dans l'évaluation des systèmes de recommandation, nous pourrions appliquer des mesures classiques de recherche d'informations pour évaluer ces moteurs : précision et rappel. Ces mesures sont largement utilisées dans les scénarios de récupération d'informations et appliquées à des domaines tels que les moteurs de recherche, qui renvoient un ensemble de meilleurs résultats pour une requête parmi de nombreux résultats possibles.

Le rappel est défini comme le nombre de documents pertinents (les instances appartenant à la catégorie pertinente) récupérés par une recherche divisé par le nombre total de documents pertinents existants. Par ailleurs, la précision est définie comme le nombre de documents pertinents (les instances appartenant à la catégorie pertinente) récupérés par une recherche divisé par le nombre total de documents récupérés par la recherche. La précision est donc la proportion de recommandations qui sont de bonnes recommandations, et le rappel est la proportion de bonnes recommandations qui apparaissent dans les principales recommandations. Lorsqu'une prédiction retourne 30 investisseurs dont seulement 20 sont pertinents (les vrais positifs) et 10 ne le sont pas (les faux positifs), mais qu'il omet 40 autres investisseurs pertinents (les faux négatifs), sa précision est de $20/30 = 2/3$ et son rappel vaut $20/(20 + 40) = 1/3$. Les mesures de rappel et précision sont utiles car elles caractérisent précisément le comportement du classifieur sur chacune des classes.
Ï

Le rappel moyen (*mean recall*) a été utilisé pour quantifier la performance des tests de ces méthodes. Cette métrique est une quantité qui répond à la question de savoir combien d'investisseurs sélectionnés sont pertinents. Il semble néanmoins dommageable de suggérer un certain nombre non-négligeable d'investisseurs qui sont de faux positifs donc non pertinents. Par conséquent, afin de minimiser la présence de ces faux positifs, il est souhaitable dans la suite du travail de considérer avec importance une autre métrique. On peut donc par la suite se baser sur l'AUC ("aire sous la courbe ROC") qui est une métrique hybride intégrant précision et rappel : la performance est synthétisée par une unique mesure et elle ne dépend pas des proportions de classe.

2 Création, pré-traitement et analyse de la base de données

Dans notre cas d'usage, la cible de la recommandation, c'est à dire l'utilisateur (*user*) dans le cas général, est ici **la transaction** (*deal*), et le "produit" recommandé (*item*) est **l'investisseur** (*investor*). Les données sont issues de la plateforme pitchbook.com qui regroupe à l'échelle globale des données de financement de marché privé des entreprises.

Le contexte du système de recommandation va jouer un rôle critique pour fournir des suggestions pertinentes. Par exemple, la situation géographique d'une transaction doit naturellement participer comme un contexte additionnel aux autres critères de taille ou de type de deal. L'information contenue dans les données doit permettre de répondre aux objectifs de la recommandation. Différents domaines tels que l'aspect temporel des données, la localité, les secteurs d'activité fournissent ainsi différents types de contextes [Charu C. Aggarwal, 2016].

Les étapes préliminaires sont :

- Constitution d'une base de données issue de pitchbook contenant des dizaines de milliers de transactions qui intègrent les caractéristiques des transactions, des entreprises et des investisseurs.
- Prétraitement de la base de données brutes (détection de données manquantes, ou doublons, homogénéisation des données, sélection des caractéristiques les plus explicatives, création de nouvelles caractéristiques pour prendre en compte la temporalité des données par exemple), encodage des variables catégorielles, vectorisation des caractéristiques.

Web Scraping – Sur la plateforme pitchbook, les types de transactions (*deals*) sont sélectionnés selon des critères déterminés. Un deal correspond à un tour de financement pour une entreprise donnée. Les pages de la plateforme pitchbook sont sauvegardées au format `.html`. Toutes les données relatives aux transactions, investisseurs, compagnies, fonds et conseillers sont respectivement stockées dans des fichiers `.json` (deals.json, investors.json, companies.json, funds.json, advisors.json). À terme, cette partie devra être automatisée afin de permettre une actualisation fréquente de la base de données et de ré-entraîner régulièrement les modèles de prédiction des investisseurs.

La base de donnée a été "scrapée" plusieurs fois. Pour notre étude, nous avons travaillé avec une base de données contenant 80,257 transactions et 26,955 investisseurs. On ne garde que les 79,786 transactions finalisées dont le statut est mentionné comme "transaction complétée" (`deal_status='Completed'`). Notons qu'aujourd'hui un scraping plus complet a été effectué avec une automatisation des tâches utilisant la librairie selenium. L'objectif est surtout de compléter la base des investisseurs avec éventuellement de nouvelles caractéristiques. Ce scraping plus complet a permis de montrer que le nombre de données manquantes est important. La figure 2.4 montre que dans les données, les *features* que nous utiliserons sont dans l'ensemble complets à 85%, exceptée la caractéristique *Verticals* complète uniquement à 28%, ce qui peut être assez problématique lorsqu'on l'utilise.

À ce jour la nouvelle base compte 469,561 transactions, 68,750 investisseurs et 340,000 entreprises. À terme, le scraping doit être fait régulièrement pour actualiser la base de données en prenant soin de ajouter un champs dans les données avec la date à laquelle les données ont été scrapées.

Critères retenus dans le modèle d'extraction des données – Comme évoqué précédemment dans la section 1, un modèle d'extraction d'un sous-ensemble des données est utile pour ajuster les recommandations à nos objectifs. Dans nos outils, nous ne retenons que les transactions supérieures à un montant de US$5M depuis le 1er janvier 2016, ainsi que les investisseurs

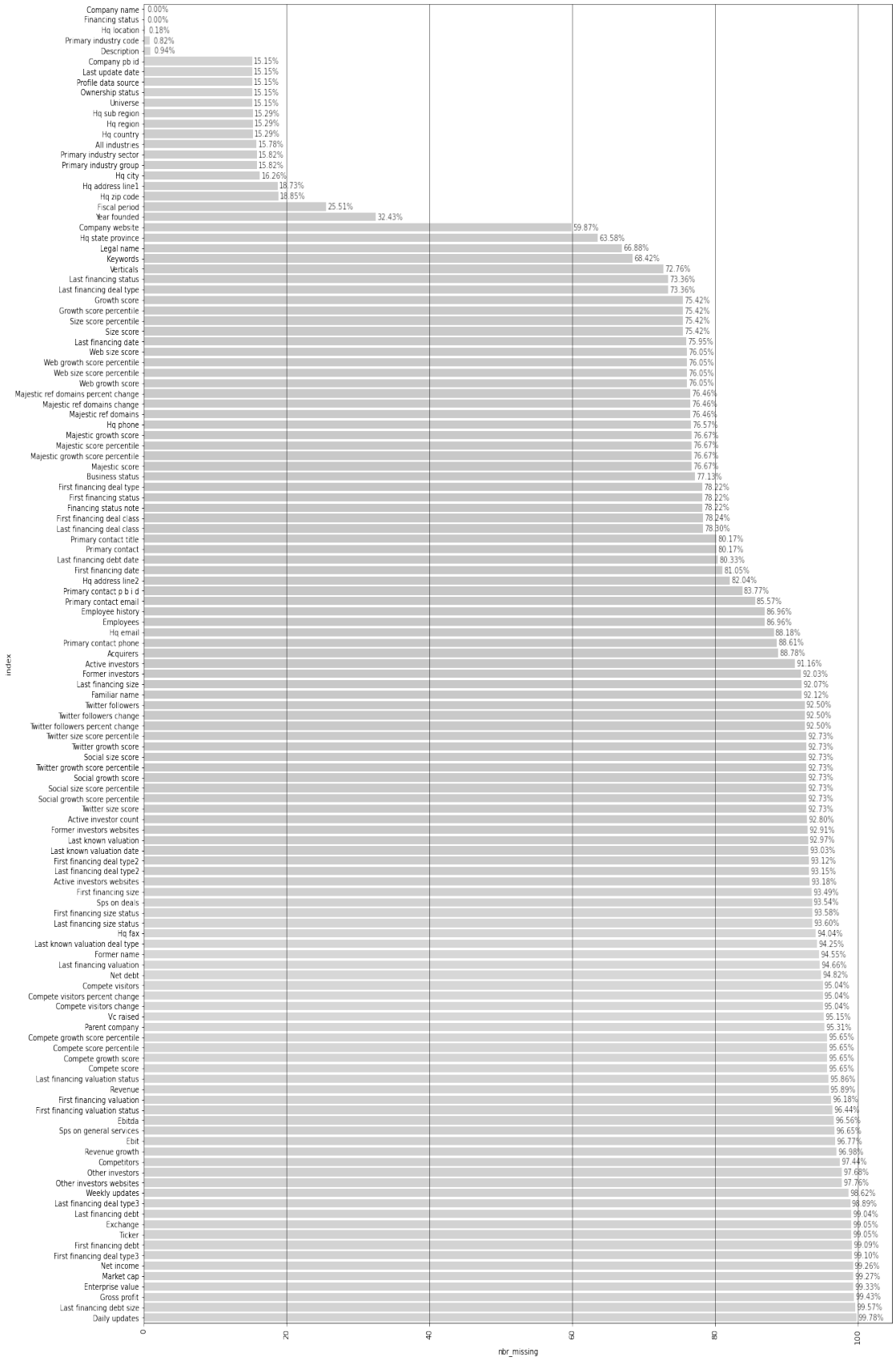

FIGURE 2.4 : Pourcentage de données manquantes pour chaque caractéristique des transactions.

```
"255": {
    "name": "Robinhood",
    "primaryIndustryGroup": "Software",
    "primaryIndustrySector": "Information Technology",
    "dealNo": 7,
    "dealDate": "31-Oct-2019",
    "dealType": "Later Stage VC",
    "Series": "Series E",
    "VCRound": "7th Round",
    "dealSize": 373.0,
    "postValuation": 7600.0,
    "impliedEV": null,
    "valuationRevenue": null,
    "ValuationEbitda": null,
    "impliedEVCashFlow": null,
    "revenue": null,
    "ebitda": null,
    "PercentAcquired": 4.91,
    "investors": "9Yards Capital, Ben Jen Holdings, Dragoneer
    Investment Group, DST Global, Friendly Hill Capital, Industrial
    Investors Group, MDT Ventures, New Enterprise Associates (Scott
    Sandell), Plus Capital, Ribbit Capital, Sequoia Capital,
    SharesPost, Thrive Capital",
    "investorCount": 13,
    "FollowOnInvestorCount": 5,
    "FollowOnInvestors": "DST Global, New Enterprise Associates,
    Ribbit Capital, Sequoia Capital, Thrive Capital",
    "firstTimeInvestorCount": 8,
    "FirstTimeInvestors": "9Yards Capital, Ben Jen Holdings, Dragoneer
    Investment Group, Friendly Hill Capital, Industrial Investors
    Group, MDT Ventures, Plus Capital, SharesPost",
    "dealSynopsis": "The company raised $373 million of Series E
    venture funding in a deal led by DST Global on October 31, 2019,
    putting the company's post-money valuation at $7.60 billion.
    Industrial Investors Group, MDT Ventures, Dragoneer Investment
    Group, Friendly Hill Capital, 9Yards Capital, Plus Capital,
    SharesPost, Ben Jen Holdings, Ribbit Capital, New Enterprise
    Associates, Thrive Capital and Sequoia Capital also participated
    in the round.",
    "dealStatus": "Completed",
    "dealFinancingStatus": "Venture Capital-Backed",
    "primaryIndustryCode": "Financial Software",
    "verticals": "FinTech, Mobile, TMT",
    "description": "Developer of an investment platform intended to
    help teach people to invest in the stock market. The company's
    platform offers commission-free trading in stocks, ETFs, options
    and cryptocurrencies, as well as margins, enabling users to invest
    in the stock market right from their phones or computers.",
    "hqLocation": "Menlo Park, CA",
    "companyWebsite": "www.robinhood.com"
},
```

```
"255": {
    "name": "Robinhood",
    "primaryIndustryGroup": "Software",
    "primaryIndustrySector": "Information Technology",
    "dealNo": 7,
    "dealDate": "31-Oct-2019",
    "dealType": "Later Stage VC",
    "Series": "Series E",
    "VCRound": "7th Round",
    "dealSize": 373.0,
    "postValuation": 7600.0,
    "impliedEV": -1,
    "valuationRevenue": -1,
    "ValuationEbitda": -1,
    "impliedEVCashFlow": -1,
    "revenue": -1,
    "ebitda": -1,
    "PercentAcquired": 4.91,
    "investors": "9Yards Capital, Ben Jen Holdings, Dragoneer
    Investment Group, DST Global, Friendly Hill Capital, Industrial
    Investors Group, MDT Ventures, New Enterprise Associates (Scott
    Sandell), Plus Capital, Ribbit Capital, Sequoia Capital,
    SharesPost, Thrive Capital",
    "investorCount": 13,
    "FollowOnInvestorCount": 5,
    "FollowOnInvestors": "DST Global, New Enterprise Associates,
    Ribbit Capital, Sequoia Capital, Thrive Capital",
    "firstTimeInvestorCount": 8,
    "FirstTimeInvestors": "9Yards Capital, Ben Jen Holdings, Dragoneer
    Investment Group, Friendly Hill Capital, Industrial Investors
    Group, MDT Ventures, Plus Capital, SharesPost",
    "dealSynopsis": "The company raised $373 million of Series E
    venture funding in a deal led by DST Global on October 31, 2019,
    putting the company's post-money valuation at $7.60 billion.
    Industrial Investors Group, MDT Ventures, Dragoneer Investment
    Group, Friendly Hill Capital, 9Yards Capital, Plus Capital,
    SharesPost, Ben Jen Holdings, Ribbit Capital, New Enterprise
    Associates, Thrive Capital and Sequoia Capital also participated
    in the round.",
    "dealStatus": "Completed",
    "dealFinancingStatus": "Venture Capital-Backed",
    "primaryIndustryCode": "Financial Software",
    "verticals": "FinTech, Mobile, TMT",
    "description": "Developer of an investment platform intended to
    help teach people to invest in the stock market. The company's
    platform offers commission-free trading in stocks, ETFs, options
    and cryptocurrencies, as well as margins, enabling users to invest
    in the stock market right from their phones or computers.",
    "hqLocation": "Menlo Park, CA",
    "companyWebsite": "www.robinhood.com",
    "hqCountryCode": "us",
    "hqCountryLat": 36.7014631,
    "hqCountryLng": -118.7559974,
    "emission": "Series E"
},
```


FIGURE 2.5 : Illustration des champs d'une transaction (**deals.json**) pour la compagnie Robinhood, capitalisée à US$7.6 milliards (`post_valuation`) dont le siège est situé à Menlo Park en Californie (`hq_location`). Cette levée a été réalisée (`deal_status : Completed`) pour un montant de US$373M (`deal_size`), et correspond à une série E (`series`), donc proche du stade final de financement privé (`deal_type`). Elle fait apparaître 8 nouveaux investisseurs (`first_time_investor_count`) parmi les 13 investisseurs (`investor_count`) ayant déjà investi dans la compagnie. Les informations sectorielles de son activité de FinTech sont inclus dans les champs `primary_industry_group`, `primary_industry_sector`, `primary_industry_code`, `verticals`. La version de droite présente des *features* additionnels.

européens, nord-américains, asiatiques et du moyen-orient. Les données incluent des dizaines de milliers de levées de fonds (*deals*) pour financer les entreprises dans leurs différents stades de développement hormis l'amorçage (ou *Seed*). Toutes les caractéristiques des transactions (*deals*) sont détaillées (taille et type de la transaction, noms et localisation des investisseurs, ainsi que leurs préférences historiques d'investissement en termes de secteur financier et géographique ou taille de transaction). Les types de transaction (*deal types*) sélectionnés sont les suivants :

- *Other Private Equity Types -> Growth/Expansion*
- *Other Private Equity Types -> PIPE*
- *All VC Stages -> Early Stage VC*
- *All VC Stages -> Later Stage VC*
- *All VC Stages -> Other Stages*
- *All Round Numbers*
- *All Series*
- *Non-Control Transactions*
- *Secondary transaction - private*
- *Public Investments -> IPO*

Plusieurs études ont été menées successivement pour essayer de quantifier les scores des modèles d'apprentissage en fonction du modèle d'extraction des données. Par exemple, nous n'avons considéré que les transactions présentant au moins deux investisseurs. Certains types de transactions ont été filtrés (`PIPE, Secondary Transaction_Private, Seed Round, Grant and Equity for Service`), mais aussi les transactions antérieures à 2018 ont aussi été supprimées dans certains cas de sous-ensembles de données. Enfin, en fonction des cas, on a aussi complètement supprimé les *deals* donc les caractéristiques considérées n'étaient pas renseignées.

Pour une transaction, il y a plusieurs investisseurs. Pour chaque *deal*, on peut accéder aux informations concernant ces investisseurs qui participent à ce tour de financement, notamment à quelle valorisation et à quels termes ils ont participé. La figure 2.5 montre un exemple de transaction avec toutes les caractéristiques retenues pour effectuer le modèle de recommandation d'investisseurs.

FIGURE 2.6 : Sélection des types d'investisseurs.

En plus des données détaillées de ces transactions, la plateforme intègre les informations concernant des dizaines de milliers d'investisseurs et d'entreprises. Nous excluons les investisseurs individuels privés tels que les business angels, ou les incubateurs. Nous sélectionnons différentes catégories d'investisseurs institutionnels (banques, assurances, fonds de pension, etc.), essentiellement les fonds d'investissement spécialisés en capital-risque (*Venture Capital*) ou en capital-investissement (*Private Equity*), mais aussi les sociétés holding d'investisseurs, les fonds souverains, les sociétés de gestion alternative (*Hedge Funds*) ou de gestion d'actifs (*Asset Manager*). Les types d'investisseurs sélectionnés sont présentés dans la figure 2.6. La figure 2.7 montre un exemple d'investisseur dans le fichier `investors.json`.

2.1 Pré-traitement des données brutes

Les programmes de pré-traitement des données brutes permettent de les découper (ou mettre en tableau), les nettoyer, les homogénéiser et les formater. On procède notamment à la détection de données dupliquées, ou manquantes.

2.1.1 Ingénierie des caractéristiques

À ce stade, on procède à de l'ingénierie des caractéristiques (*feature engineering*) qui consistent à la création de nouvelles caractéristiques et la sélection des caractéristiques les plus significatives. Les nouvelles caractéristiques ont pour but de capturer des informations supplémentaires qui ne sont pas directement accessibles dans l'ensemble des caractéristiques d'origine. Ces nouvelles caractéristiques vont permettre d'améliorer les performances de prédiction des modèles d'apprentissage.

- Ingénierie des caractéristiques : processus de création de nouvelles caractéristiques à partir de données brutes pour augmenter la puissance prédictive de l'algorithme d'apprentissage. Les caractéristiques conçues doivent capturer des informations supplémentaires qui ne sont pas facilement visibles dans l'ensemble de caractéristiques d'origine.
- Sélection de caractéristiques : Processus de sélection du sous-ensemble clé de caractéristiques pour réduire la dimensionnalité du problème d'apprentissage.

L'ingénierie et la sélection de caractéristiques augmentent l'efficacité du processus d'apprentissage qui tend à extraire les informations essentielles contenues dans les données. Ces processus améliorent également les performances de ces modèles pour classifier les données d'entrée avec précision et prédire les résultats pertinents de façon plus consistante. L'ingénierie et la sélection de caractéristiques peuvent également être combinées afin de faciliter l'apprentissage. Cela se fait grâce à l'amélioration puis à la réduction du nombre de caractéristiques nécessaires à l'étalonnage ou l'apprentissage d'un modèle. Les caractéristiques sélectionnées sont un ensemble minimum de variables indépendantes qui expliquent les modèles des données, et prédisent correctement des résultats. Il n'est pas toujours nécessaire d'effectuer une ingénierie de caractéristiques ou une sélection de caractéristiques. Cela dépend des données, de l'algorithme sélectionné et de l'objectif de l'expérience.

Dans notre cas, les nouvelles caractéristiques sont basées sur des aspects de temporalités des données, ou bien plus simplement, il peut s'agir de la conversion d'une variable géographique de type (ville, état, pays) en région géographique plus étendue telle que EMEA (Europe Middle East & Africa), APAC (Asia Pacific), etc. Il peut également s'agir, encore plus simplement, d'homogénéiser ces données géographiques. À ce stade, nous n'utilisons que les données issues de Pitchbook, mais à terme, avec l'utilisation d'autres sources de données (*e.g.* Ipreo, IHS Markit, Rothschild, données internes de Praexo), il sera important d'homogénéiser, entre autres, la terminologie décrivant les secteurs d'activités correspondant aux transactions.

```
"32": {
    "name": "Warburg Pincus",
    "description": "Founded in 1966, Warburg Pincus is a private equity firm based in New York, New York. The firm invests in companies based in the
    Americas, Europe, Middle East, and Asia. The firm seeks to invest in the like energy, financial services, healthcare and consumer, industrial and
    business services, technology, real estate, media, and telecommunication sectors. ",
    "primaryInvestorType": "PE/Buyout",
    "capitalUnderManagement": 58000.0,
    "lastUpdateDate": "14-Apr-2020",
    "yearFounded": 1966,
    "hqLocation": "New York, NY",
    "investments": 113,
    "activeInvestedCompanies": 87,
    "campanyName": "Privitar",
    "dealDate": "06-Apr-2020",
    "dealSize": 80.0,
    "preferredIndustry": "Commercial Products, Commercial Services, Communications and Networking, Consumer Durables, Healthcare Services, Media,
    Software, Transportation",
    "preferredInvestmentTypes": "Buyout/LBO, Early Stage VC, Later Stage VC, PE Growth/Expansion, Seed Round",
    "preferredCompanyValuation": "100.00 - 15,000.00",
    "preferredDealSize": "50.00 - 500.00",
    "preferredInvestmentAmount": "20.00 - 1,000.00",
    "preferredOtherPreferences": "Long-Term Investor, Prefers minority stake, Seeks ESG investments, Will syndicate",
    "preferredGeography": "Brazil, Canada, China, Europe, Hong Kong, India, Middle East, Singapore, Taiwan, United States, Vietnam",
    "preferredVerticals": "AdTech, Advanced Manufacturing, CleanTech, Digital Health, FinTech, Gaming, HealthTech, Industrials, Infrastructure, Life
    Sciences, LOHAS & Wellness, Manufacturing, Real Estate Technology, SaaS, TMT",
    "postValuation": 400.0,
    "investmentsInTheLast12Months": 17,
    "keywords": "business model, hedging fund, venture capital",
    "verticals": "",
    "primaryIndustryGroup": "Capital Markets/Institutions",
    "primaryIndustrySector": "Financial Services",
    "investorWebsite": "www.warburgpincus.com",
    "primaryContact": "Mark M. Colodny",
    "primaryContactTitle": "Managing Director, Executive Management, Technology, Media and Telecommunications",
    "primaryContactPhone": "+1 (212) 878-0600",
    "primaryContactEmail": "mcolodny@warburgpincus.com"
},
```

```
"32": {
    "name": "Warburg Pincus",
    "description": "Founded in 1966, Warburg Pincus is a private equity firm based in New York, New York. The firm invests in companies based in the
    Americas, Europe, Middle East, and Asia. The firm seeks to invest in the like energy, financial services, healthcare and consumer, industrial
    and business services, technology, real estate, media, and telecommunication sectors. ",
    "primaryInvestorType": "PE/Buyout",
    "capitalUnderManagement": 58000.0,
    "lastUpdateDate": "14-Apr-2020",
    "yearFounded": 1966,
    "hqLocation": "New York, NY",
    "investments": 113,
    "activeInvestedCompanies": 87,
    "campanyName": "Privitar",
    "dealDate": "06-Apr-2020",
    "dealSize": 80.0,
    "preferredIndustry": "Commercial Products, Commercial Services, Communications and Networking, Consumer Durables, Healthcare Services, Media,
    Software, Transportation",
    "preferredInvestmentTypes": "Buyout/LBO, Early Stage VC, Later Stage VC, PE Growth/Expansion, Seed Round",
    "preferredCompanyValuation": "100.00 - 15,000.00",
    "preferredDealSize": "50.00 - 500.00",
    "preferredInvestmentAmount": "20.00 - 1,000.00",
    "preferredOtherPreferences": "Long-Term Investor, Prefers minority stake, Seeks ESG investments, Will syndicate",
    "preferredGeography": "Brazil, Canada, China, Europe, Hong Kong, India, Middle East, Singapore, Taiwan, United States, Vietnam",
    "preferredVerticals": "AdTech, Advanced Manufacturing, CleanTech, Digital Health, FinTech, Gaming, HealthTech, Industrials, Infrastructure, Life
    Sciences, LOHAS & Wellness, Manufacturing, Real Estate Technology, SaaS, TMT",
    "postValuation": 400.0,
    "investmentsInTheLast12Months": 17,
    "keywords": "business model, hedging fund, venture capital",
    "verticals": "",
    "primaryIndustryGroup": "Capital Markets/Institutions",
    "primaryIndustrySector": "Financial Services",
    "investorWebsite": "www.warburgpincus.com",
    "primaryContact": "Mark M. Colodny",
    "primaryContactTitle": "Managing Director, Executive Management, Technology, Media and Telecommunications",
    "primaryContactPhone": "+1 (212) 878-0600",
    "primaryContactEmail": "mcolodny@warburgpincus.com",
    "hqCountryCode": "us",
    "hqCountryLat": 43.1561681,
    "hqCountryLng": -75.8449946,
    "histIndustryGroup": "Software, Communications and Networking, Commercial Services, Commercial Transportation, Services (Non-Financial),
    Commercial Products, Retail, Transportation, Healthcare Devices and Supplies, Other Financial Services, Media, IT Services, Computer Hardware,
    Pharmaceuticals and Biotechnology, Healthcare Technology Systems, Exploration, Production and Refining",
    "histVerticals": "FinTech, SaaS, TMT, Mobile, Industrials, Mobility Tech, Ridesharing, Supply Chain Tech, Artificial Intelligence & Machine
    Learning, Autonomous cars, Real Estate Technology, Big Data, EdTech, E-Commerce, Advanced Manufacturing, Cybersecurity, Micro-Mobility, \u00a0,
    Beauty, Impact Investing, Mobile Commerce, InsurTech, Marketing Tech, Robotics and Drones, Digital Health, Life Sciences, HealthTech, Oil & Gas",
    "histHqCountryCode": "cn, za, id, it, us, in, gb, sg, ca",
    "histDealSize": "14000.0, 1438.71, 1500.0, 600.0, 613.28, 300.0, 673.6, 335.0, 550.0, 130.0, 69.7, 100.0, 96.0, 215.0, 60.0, 132.0, 65.0, 24.45,
    80.0, 76.5, 108.0, 220.0, 47.5, 45.0, 25.18, 35.0, 1102.0, 20.0, 25.0, 142.68, 255.0, 46.89, 19.39, 43.77, 150.0, 200.0, 76.0, 120.0, 74.0,
    44.0, 47.0, 180.0, 140.0, 500.0",
    "histIndustrySector": "Information Technology, Business Products and Services (B2B), Consumer Products and Services (B2C), Healthcare, Financial
    Services, Energy",
    "histPrimaryIndustryCode": "Financial Software, Telecommunications Service Providers, Logistics, Road, Real Estate Services (B2C), Educational
    Software, Other Commercial Products, Social/Platform Software, Internet Retail, Network Management Software, Automation/Workflow Software, Other
    Transportation, Therapeutic Devices, Business/Productivity Software, Consumer Finance, Other Commercial Services, Information Services (B2C), IT
    Consulting and Outsourcing, Application Software, Other Hardware, Biotechnology, Medical Records Systems, Construction and Engineering, Energy
    Production"
},
```


FIGURE 2.7 : Exemple d'un investisseur dans le fichier `investors.json` avant et après un prétraitement.

Parmi la création de nouvelle caractéristique, on crée notamment, pour chaque investisseur, les suivantes :

- Historique des secteurs d'activité des transactions (*investors_histIndustrySector*, *investors_histIndustryGroup*, *investors_histPrimaryIndustryCode*, *investors_histVerticals*)
- Historique des zones géographiques des transactions (*investors_histHqCountryCode*)
- Historique de la taille des transactions (*investors_histDealSize*).

La figure 2.7 illustre la création des nouvelles *features* relatives à l'historique des transactions d'un investisseur.

La création de nouvelles caractéristiques (fig. 2.5) peut aussi permettre de simplifier la description des transactions en prenant en considération les redondances dans les dénominations. Ainsi, le type (*deal_dealType*) et la série (*deal_Series*) des transactions peuvent être fusionnés en une nouvelle caractéristique nommée l'émission de la transaction (*deal_emission*). La terminologie de la série de la transaction est tout d'abord simplifiée en ne prenant que la première lettre : par exemple, on regroupe ainsi les séries AA, A1, A2, A3, 1 en un unique terme désigné série A. Ensuite, pour encore simplifier, les types de transactions désignées *Early Stage VC* et *Later Stage VC* sont respectivement associés à une émission de transaction nommée Série A et Série F.

La nouvelle caractéristique de stade d'émission de la transaction (*deal_emission*) est ensuite utilisée dans un pré-conditionnement des résultats de recommandation. Dans une nomenclature spécifique (fig. 2.8), on prend en compte un recoupement des univers des investisseurs qui définit une certaine perméabilité en termes d'émission et de taille de transaction. Un filtre statique peut être appliqué sur les investisseurs suggérés par le modèle en astreignant la recommandation de manière à ce qu'elle corresponde à un cas acceptable conformément à cette nomenclature. Cette zone d'acceptation est délimitée par la zone grisée sur la figure 2.8. Par exemple, la zone grisée des deux premières lignes de la figure 2.8 montrent que pour un deal d'une Série A ou B, on élargira la possibilité d'avoir un investisseur suggéré aux situations où la série est une Série A, B, C ou D, et où le type de deal est *Seed* ou *Corporate*, et si en plus la taille du deal est supérieure à $100M, on élargira aux types de deal PIPE et IPO.

Après avoir enrichi le jeu de données d'apprentissage, une sélection de caractéristiques les plus pertinentes est nécessaire pour avoir de bonnes performances. La sélection de caractéristiques est alors effectuée pour éliminer celles qui sont non-pertinentes, redondantes ou fortement corrélées. À partir de données brutes semi-structurées au format `.json`, la constitution d'un tableau de données au format `.csv` peut être éventuellement effectuée pour faciliter la manipulation des données avant d'entraîner un modèle de machine learning.

La plupart des algorithmes de *machine learning* requièrent des données d'entrée sous la forme de matrice numérique où chaque ligne est un échantillon et chaque colonne une caractéristique. Pour chaque investisseur, il y plusieurs caractéristiques contenant chacune de multiples champs (*one-to-many entity-relationship*) : les préférences et l'historique des investisseurs en terme de type de transactions, et de secteurs d'activité par exemple. Dans ce processus, on procède donc à la "mise à plat" des données : chaque ligne du tableau correspondant à une transaction est dupliquée en fonction du nombre d'éléments de chaque variable catégorielles. On convertit les relation *one-to-many* en relations *one-to-one*. Par exemple, si nous créons la caractéristique *investors_histPrimaryIndustryCode* qui traduit pour chaque investisseur l'historique d'un descriptif sectoriel d'activité spécifique (*PrimaryIndustryCode*) pour toutes ses transactions, nous obtenons une liste de plusieurs éléments qui devra être séparées pendant la mise en tableau et aura pour conséquence de multiplier les lignes associées à une transaction.

FIGURE 2.8 : Nomenclature définissant le recoupement entre le stade d'émission et la taille de la transaction.

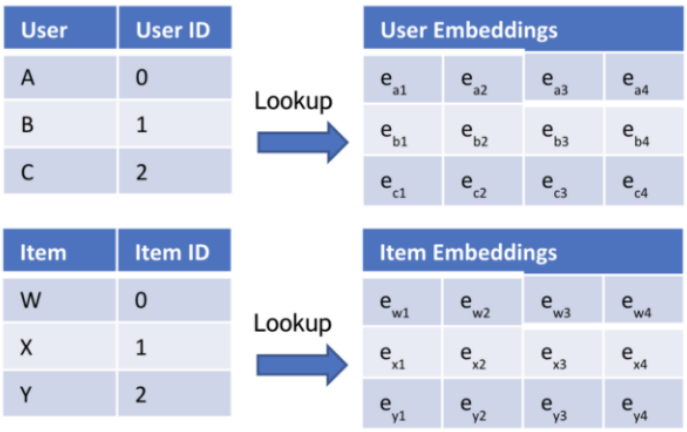

FIGURE 2.9 : Illustration d'un *embedding* de dimension 4.

2.1.2 Encodage des caractéristiques catégorielles en variables numériques

Les différentes variables catégorielles sont séparées lorsqu'elles sont des champs multiples, puis encodées numériquement selon la méthode d'encodage de labels (*label encoding*) c'est à dire qu'un nombre aléatoire est assigné à chaque catégorie unique. Le choix d'effectuer un *label encoding* au lieu d'un *one-hot encoding* peut être discutable. En effet, contrairement au *one-hot encoding*, le *label encoding* peut engendrer une hiérarchie artificielle entre les catégories. En effet, les algorithmes d'apprentissage peuvent mal interpréter les valeurs numériques correspondant aux différentes catégories en les traitant avec un certain ordre de priorité. Cela dit, il est, d'une part, plus pratique d'utiliser le *label encoding* puisqu'il a l'avantage de ne pas multiplier excessivement le nombre de *features* comme le fait le *one-hot encoding*, et d'autre part, l'implémentation est telle que les valeurs numériques d'assignation sont tirées aléatoirement ce qui réduit le problème de hiérarchie des catégories. Le problème de cet encodage est qu'il limite la captation des relations entre les catégories. Il s'agit là d'un sujet très important qui détermine la qualité du système de recommandation. L'encodage par étiquettes est réalisé à l'aide de la librairie *NumericalEncoder* de aikit qui optimise le process en tenant compte, entre autres, des valeurs manquantes. Il permet aussi de fusionner des données dans certains cas avec trop peu d'observations pour réduire la dimension des résultats et éviter le sur-apprentissage. Une simple normalisation est enfin effectuée pour ramener simplement toutes les valeurs entre 0 et 1 (*MinMaxScaler*).

2.1.3 Vectorisation des données (*i.e. Embedding*)

On rappelle que l'*embedding* consiste en une représentation distribuée des données. L'espace vectoriel obtenu fournit une projection des catégories, permettant aux catégories proches ou liées de se regrouper naturellement. En section 3.2, on montrera que les modèles développés sont souvent basés sur un réseau de neurones à propagation avant (*feedforward neural network*), on entraîne le réseau de neurones à apprendre la vectorisation des catégories d'un espace à très haute dimension vers un espace à dimension moindre (*entity embedding*) [Guo and Berkhahn, 2016]. Chaque catégorie encodée est mappée sur un vecteur distinct, et les propriétés du vecteur sont adaptées ou apprises grâce au réseau de neurones. Le remplacement de chaque valeur de catégorie par son vecteur d'*embedding* est une étape majeure (fig. 2.9). Il semble donc important d'approfondir les méthodes employées concernant cette étape afin de capter au mieux l'information contenue dans les variables catégorielles à cardinalité élevée. Ce sujet fait l'objet de recherches importantes [Cerda et al., 2018]. L'utilisation de techniques de NLP telles que les transformers par exemple peuvent être utilisées dans pour des recommandations séquentielles [Zhao et al., 2020a].

2.2 Analyse statistiques des données

Afin de mieux appréhender l'ensemble des données, on procède à une analyse statistique qui peut permettre de déceler des problèmes de données manquantes ou autres, ou bien tout simplement d'identifier des tendances intéressantes à explorer. Dans cette section, une analyse non-exhaustive des données est présentée. Par exemple, la figure 2.10 montre que, pour une transaction en moyenne, le nombre d'investisseurs est de l'ordre de 8 investisseurs et que, parmi ces 8 investisseurs, la moitié était déjà actionnaire de l'entreprise, et l'autre moitié se compose de nouveaux investisseurs.

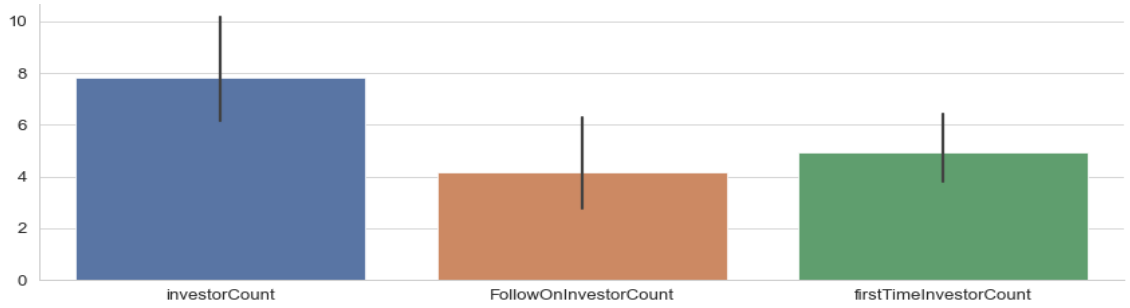

FIGURE 2.10 : Pour une transaction, nombre moyen d'investisseurs (*à gauche*), d'investisseurs historiques (*au milieu*), et de nouveaux investisseurs (*à droite*)

Les séries B sont généralement menées par les mêmes acteurs que le tour précédent avec comme différence l'addition de nouveaux fonds de capital-risque spécialisés dans des financements d'entreprises plus matures.

Series	investorCount	FollowOnInvestorCount	firstTimeInvestorCount	FollowOnInvestorCount_percent	firstTimeInvestorCount_percent
A	4.90	2.30	4.04	43.04	84.15
B	5.14	2.69	3.18	50.91	62.98
C	5.53	3.15	3.13	55.53	57.75
D	5.76	3.52	3.13	57.24	57.40
E	6.06	3.84	3.32	59.64	57.67
F	6.80	3.90	3.89	59.97	54.50
G	6.60	3.33	4.34	54.12	62.19
H	8.40	4.23	5.85	58.55	61.55
I	9.00	4.29	6.00	37.07	74.05
J	16.75	13.67	6.50	46.89	64.83
K	11.33	1.00	11.00	10.00	96.67

FIGURE 2.11 : Par série, nombre moyen d'investisseurs, d'investisseurs historiques, et de nouveaux investisseurs.

Si on fait cette analyse par série, la tableau de la figure 2.11 montre que cette tendance est bien confirmée pour les premiers tours de financement. En revanche, la série K présente un comportement particulier puisqu'il ne s'agit que de nouveaux entrants.

3 Création d'un système de recommandation d'investisseurs

3.1 Contexte

Un filtrage statique est directement accessible sur la plateforme pitchbook.com. Il est alors possible d'obtenir une liste d'investisseurs correspondant aux transactions passées en fonction de leurs différentes caractéristiques. Afin de dépasser le filtrage statique des investisseurs potentiellement intéressés par l'investissement dans une entreprise, les équipes de Praexo s'orientent vers un filtrage intelligent entraîné via des données intégrant par exemple des aspects temporels. En particulier, on peut favoriser un investisseur ayant investi il y a moins d'un an dans le même secteur d'industrie ou dans la même zone géographique. Après avoir entraîné un modèle d'apprentissage statistique qui va regrouper les investisseurs selon leurs caractéristiques, on peut alors bénéficier d'une certaine perméabilité entre les différents groupes pour extraire des suggestions d'investisseurs pertinents. À terme, après intégration à l'interface de Praexo, l'outil de recommandation permettrait de suggérer une liste d'investisseurs à une entreprise dont certaines propriétés sont connues et qui cherche à se financer. Ces mêmes investisseurs seraient alors prévenus par la volonté de l'entreprise de lever des fonds.

Le besoin de sérendipité – Tandis qu'il apparaît plus facile de rester dans une certaine zone de confort en ne suggérant que des investisseurs prévisibles par l'utilisateur, l'objectif d'un système de recommandation réside en particulier dans le fait de remplir un besoin de sérendipité, c'est à dire de surprendre l'utilisateur avec des résultats à la fois insolites en quelques sortes et en même temps très pertinents pour le cas d'usage sélectionné. Ainsi, le système de recommandation doit pouvoir suggérer des investisseurs auxquels personne n'aurait pensé mais qui s'avèrent être des solutions intéressantes et judicieuses pour la transaction. En effet, dans la pratique, les conseillers des banques d'affaires ont généralement tendance à contacter en premier lieu les investisseurs qu'ils connaissent et qui appartiennent à leur réseau habituel.

Dans une phase de prospection de faisabilité et d'évaluation des performances, trois approches totalement différentes ont été réalisées et testées pour le système de recommandation d'investisseurs. Les différentes approches présentent des points communs concernant la préparation des données, même si cela peut différer légèrement dans certains cas. Pour pouvoir mieux comparer les approches, un effort a été fait, dans l'ensemble, pour démarrer sur une même base de départ, autant que possible : cela inclue le scrapping des données initiales, ainsi que le pré-traitement des données avec la création de nouvelles caractéristiques (homogénéisation des critères géographiques, création de variables prenant en compte un aspect temporel, c'est à dire historique des caractéristiques des transactions, modèle d'extraction des caractéristiques significatives et bien renseignées dans la base initiale) mais aussi l'encodage des variables catégorielles et la standardisation des variables.

3.2 Première approche

L'algorithme de suggestion procède en trois phases : en construisant d'abord un réseau de neurones associant les caractéristiques d'une entreprise désirant se financer aux caractéristiques d'un investisseur, puis en construisant une mesure de distance entre investisseur et investisseur idéal, enfin en sélectionnant les investisseurs qui sont à une distance inférieure à seuil donné.

Le réseau de neurones utilisé est un auto-encodeur (*DAE i.e. Deep Auto-Encoder*). Il est bien connu qu'un auto-encodeur a la capacité inhérente d'apprendre une représentation compacte des données. Les auto-encodeurs ont été largement utilisés en raison de ses performances dans la réduction de dimension des données, mais aussi pour le débruitage, l'extraction des *features* et la reconstruction des données. Des études très récentes [Ferreira et al., 2020 ; Bobadilla et al., 2020 ; Haghighi et al., 2020 ; Zhu et al., 2019] ont montré que l'utilisation d'un auto-encodeur révèle de

meilleures performances pour un système de recommandation que la SVD qui sera l'objet de notre deuxième approche en section 3.3.

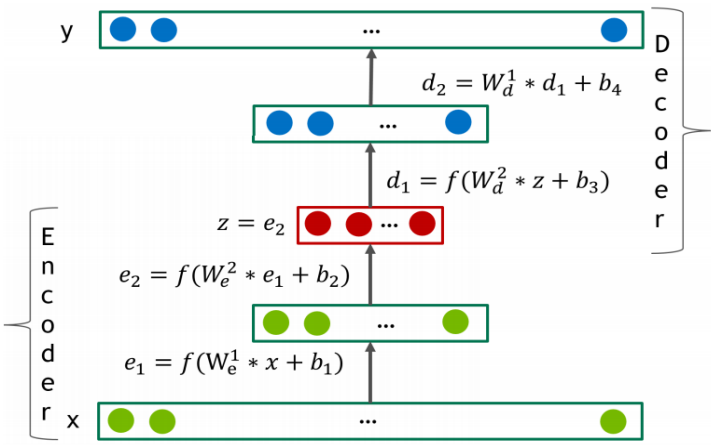

FIGURE 2.12 : Architecture de l'auto-encodeur.

Comme le montre la figure 2.12, l'auto-encodeur se compose de deux parties : un encodage et un décodage. L'encodage est un réseau de neurones à propagation avant entièrement connecté (*fully connected feedforward neural network*) qui va lire les entrées et compresser des données sous la forme :

$$z = \texttt{encode}(x), \tag{2.1}$$

puis le décodage va lire cette représentation compactée z, et recomposer les données en sortie y, à la même dimension que les données d'entrée, soit :

$$y = \texttt{decode}(z) = \texttt{decode}(\texttt{encode}(x)). \tag{2.2}$$

L'erreur est calculée en prenant la différence entre les entrées d'origine x et le signal reconstruit y, soit $e = x - r$. Le réseau auto-encodeur apprend à réduire cette erreur.

Pour une expérimentation rapide, on utilise les librairies *deep learning* de Keras s'exécutant sur TensorFlow. L'un des avantages principaux d'utiliser Keras est de pouvoir développer rapidement, et en particulier, tester des architectures de modèles exotiques. De nombreux tests d'architecture ont été réalisés afin d'optimiser les performances du système de recommandation. En augmentant le nombre de couches d'un MLP (*Multi-Layer Perceptron*), on augmente les performances du réseau de plus en plus profond, mais les problèmes de gradients évanescents peuvent alors apparaître. Pour pallier ce problème, on utilise des blocs de couches denses (*fully-connected*) avec une fonction d'activation ReLU. On intègre, dans l'architecture des huit couches cachées du réseau, une couche de *dropout* dont on rappelle schématiquement le principe dans la figure 2.13. Cette couche permet d'éviter le sur-apprentissage. Le modèle est alors compilé en utilisant l'optimisation Adam [Kingma and Ba, 2015] avec comme fonction de coût l'erreur quadratique moyenne (*MSE i.e. Mean Squared Error*), et comme métrique l'erreur absolue moyenne (*MAE i.e. Mean Absolute Error*). Adam combine respectivement les bénéfices des algorithmes AdaGrad et RMSProp pour fournir un algorithme d'optimisation capable de gérer des gradients clairsemés sur des problèmes de bruit.

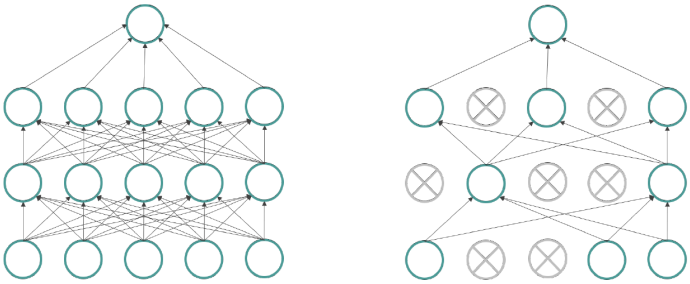

FIGURE 2.13 : Principe du *dropout* : on ignore de manière aléatoire certains neurones.

Concernant le choix des caractéristiques, plusieurs configurations ont été testées. Les résultats sont tous assez proches en terme de précision. Typiquement, on prend par exemple 7 caractéristiques d'entrée pour les transactions :

- `deals_primaryIndustryGroup`,
- `deals_primaryIndustrySector`,
- `deals_primaryIndustryCode`,
- `deals_emission`,
- `deals_dealSize`,
- `deals_hqCountryCode`,
- `deals_FollowOnInvestors`,

et 5 caractéristiques d'entrée pour les investisseurs :

- `investors_primaryInvestorType`,
- `investors_histIndustryGroup`,
- `investors_histHqCountryCode`,
- `investors_histPrimaryIndustryCode`,
- `investors_hqCountryCode`.

Comme évoqué dans les sections 2.1.2 et 2.1.3, on procède à un encodage par label puis un *embedding* des données catégorielles.

Dans cette méthode, on entraîne un modèle pour attribuer des poids aux caractéristiques des investisseurs [Chollet, 2017]. On entraîne le modèle avec 80% des données encodées, et normalisées (MinMaxScaler) et on le teste avec 20% restants. Les coefficients de caractéristiques, une fois prédits, vont permettre ensuite, par une mesure de distance dans un sous-espace (*embedding*), de trouver les investisseurs les plus proches en termes de caractéristiques. Plusieurs méthodes pour le calcul de la distance entre les paramètres estimés par le modèle et les investisseurs suggérés ont été testées.

Pour résumer, la suggestion se fait donc en deux étapes :

- Prédiction des caractéristiques des investisseurs en sortie de réseau, en fonction des caractéristiques de la transaction.
- Suggestion des investisseurs dont les caractéristiques sont les plus proches de celles prédites par le modèle d'apprentissage.

Le but principal de l'intégration des catégories (*embedding categories*) est de cartographier des catégories similaires proches les unes des autres dans l'espace d'intégration (*embedding space*) [Guo and Berkhahn, 2016]. La vectorisation des catégories permet de réduire la dimension d'un espace de très-haute dimension (*entity embedding*) et facilite le calcul de la distance entre les

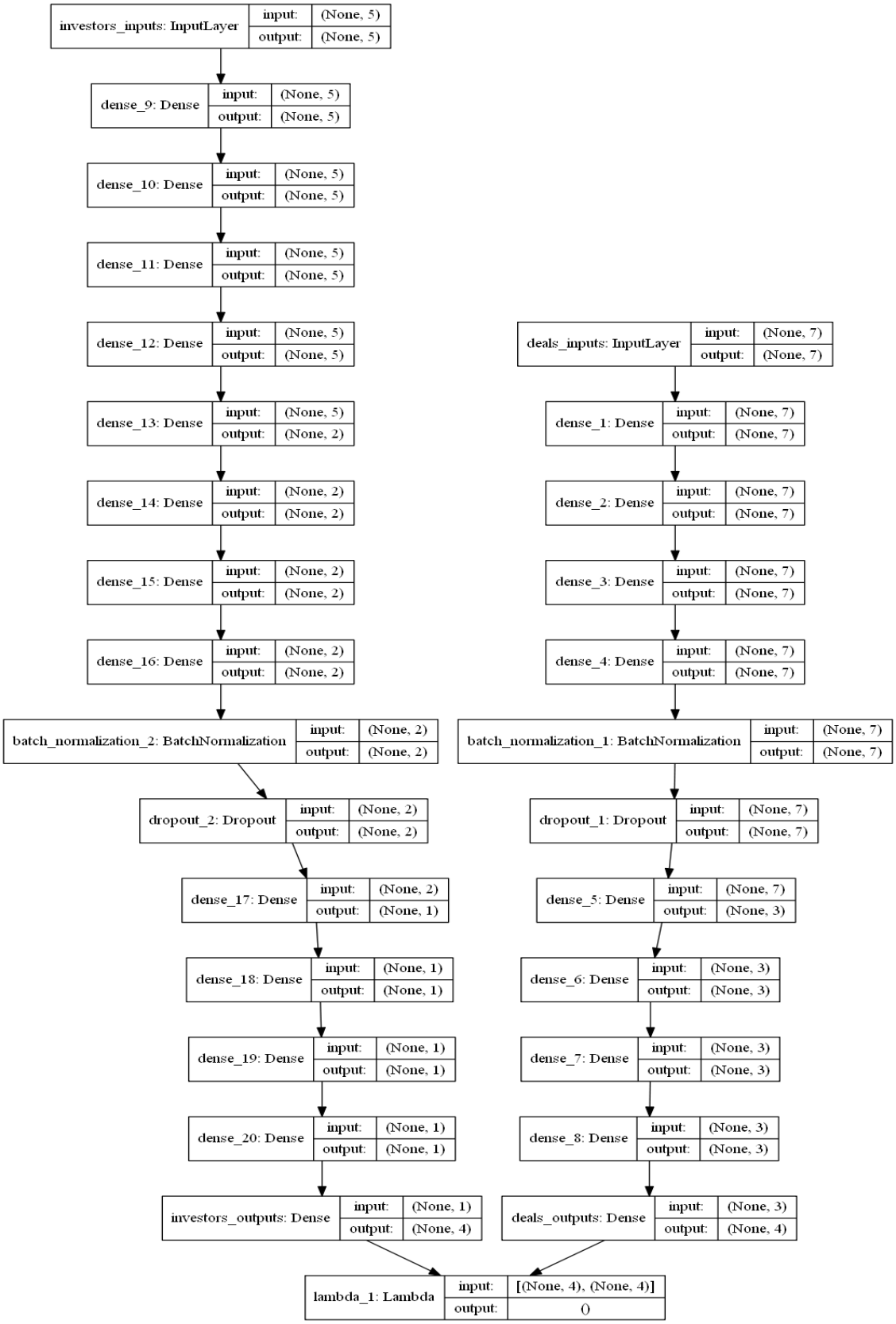

FIGURE 2.14 : Architecture du réseau de neurones utilisé dans la première approche (7 *features* pour les transactions et 5 pour les investisseurs).

caractéristiques prédites et les caractéristiques des investisseurs [Guo and Berkhahn, 2016]. Ainsi, la similarité est calculée telle que :

```
distance = np.linalg.norm(deal_embedding - investor_embedding).
```

Notons que les données catégorielles non-organisées conduisent souvent à une plus grande cardinalité des variables catégorielles et donnent lieu à plusieurs problèmes lors de l'utilisation de l'encodage.Un premier défi est que l'ensemble de données peut contenir des représentations morphologiques différentes de la même catégorie. Par exemple, pour une variable catégorielle d'un nom de société, il n'est pas évident que "Pfizer International LLC", "Pfizer Limited" et "Pfizer Korea" soient des noms différents pour la même entité, mais ils sont certainement liés. Ici, nous nous appuyons sur l'intuition que ces entités sont probablement proches dans l'espace des caractéristiques et non pas des catégories indépendantes. Dans les données brutes, des erreurs telles que des fautes de frappe peuvent entraîner des variations morphologiques des catégories. Sans nettoyage complet et homogénéisation des données, différentes représentations de chaînes de caractères de la même catégorie vont conduire à des vecteurs codés complètement différents. Un autre défi connexe est celui des catégories encodées qui n'apparaissent pas dans l'ensemble d'apprentissage [Cerda et al., 2018].

3.3 Deuxième approche

La deuxième approche, plus rudimentaire, est basée sur une technique de filtrage collaboratif (*collaborative filtering*) puisqu'elle ne prend en compte que les caractéristiques de la transaction et non celles des investisseurs mis à part leur nom. Comme présentée dans la section 1.1, cette technique va fournir des prédictions sur les préférences d'investisseurs associées à une transaction à partir de la collecte des préférences de nombreuses transactions. On se base sur l'hypothèse que si les transactions (utilisateurs) ont un sous-ensemble de "préférences similaires", ils sont plus susceptibles d'avoir également des "préférences similaires" sur d'autres investisseurs (éléments) invisibles.

Pour réaliser cette deuxième approche, on procède en un encodage *one-hot* des caractéristiques de transaction pour générer une représentation parcimonieuse de la connexion des investisseurs des transactions aux investisseurs par une matrice creuse (*sparse matrix*) de taille [*nombre_d_investisseurs_dans_les_deals*, *nombre_d_investisseurs_uniques*]. Cette matrice est appelée matrice de notation (*rating matrix*) puisque chaque élément de cette matrice correspond au nombre de transactions auxquelles chaque investisseur a participé. Cette matrice de notation est une matrice creuse puisque le nombre de valeurs non-nulles est de 93463 parmi 247148910, le nombre total de valeurs. La parcimonie importante de ces données est donc caractérisée par un taux d'occupation de 0.04%. Une analyse en composantes principales (PCA) permet alors de compresser (*SVD for a Low-Dimensional Embedding*) cette matrice contenant l'information qui sera ensuite utilisée pour émettre des suggestions d'investisseurs [Shi et al., 2019]. Pour ce faire, deux types de modèles d'apprentissage ont été utilisés dans cette approche : un réseau de neurones et un algorithme d'arbre de boosting de gradient (implémentation *Scikit-learn's Gradient Boosted Decision Trees (GBDT)*). La méthode de gradient boosting est une méthode d'ensemble. L'algorithme GradientBoosting de scikit learn est une implémentation performante de l'algorithme d'arbre de décision à gradient boosté. Il est utilisé pour le classement, la classification et d'autres tâches d'apprentissage automatique. XGboost est une implémentation de GBDT avec randomisation (il utilise l'échantillonnage de colonnes et l'échantillonnage de lignes). L'échantillonnage de lignes est possible en n'utilisant pas toutes les données d'apprentissage pour chaque modèle de base du GBDT. Au lieu d'utiliser toutes les données d'entraînement pour chaque modèle de base, nous échantillonnons un sous-ensemble de lignes et n'utilisons que ces lignes de données pour créer chacun des modèles de base. Cela garantit qu'il y a moins de risques de *over-fitting*, ce qui est un problème majeur avec le simple GBDT que XGBoost tente de résoudre en utilisant cette randomisation.

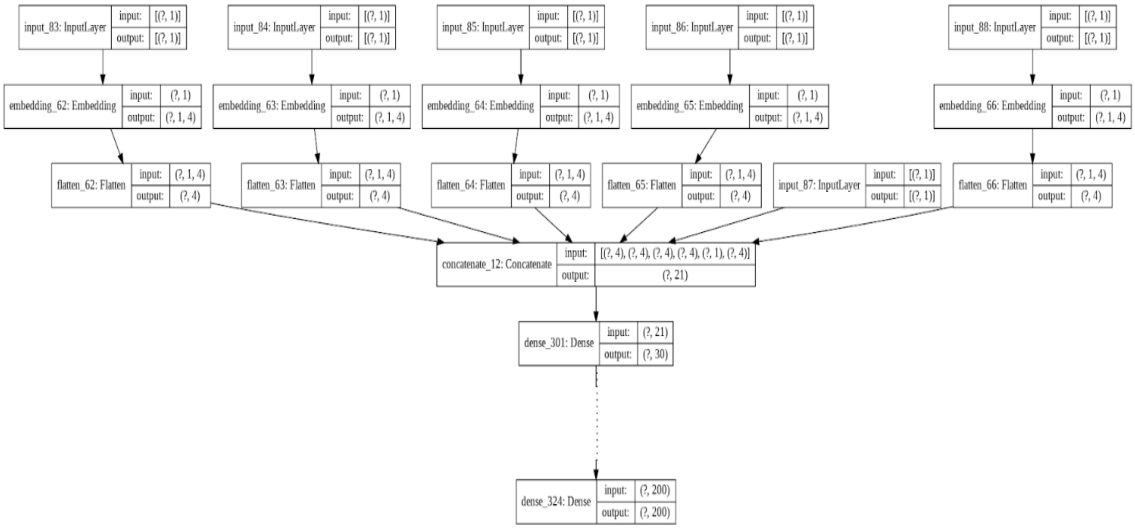

FIGURE 2.15 : Architecture du réseau de neurones utilisé dans la deuxième approche (5 *features* catégorielle et 1 numérique).

Par exemple, on utilise en entrée du réseau, les 6 *features* de transaction suivants :

- `deals_primaryIndustryGroup`,
- `deals_primaryIndustrySector`,
- `deals_verticals`,
- `deals_dealType`,
- `deals_dealSize`,
- `deals_hqLocation`

qui sont toutes catégorielles sauf `deals_dealSize`, donc elles sont encapsulées dans un sous-espace comme l'illustre les couches *embedding* la figure 2.15. La sortie du réseau est le *feature* `deals_investors`.

Le tableau de la figure 2.16 compare les résultats de précision avec cette approche en utilisant les deux types de modèles pour effectuer la prédiction : le réseau de neurones et l'algorithme XGBoost. Les deux modèles ont donné des résultats de performance très similaires.

- 2 différents modèles :
  - Gradient Boosting Regressor Algorithm (XGBoost)
  - Feed Forward Neural Network (FFNN)

4 tests différents :		Résultats
FFNN	(sortie : liste de 100 prédictions)	4525 intersections non-nulles sur 10575 (42.8%) => 25.8 % rappel total
FFNN	(sortie : liste de 1000 prédictions)	8173 intersections non-nulles sur 10575 (77.3%) => 44.7 % rappel total
XGBoost (sortie : liste de 100 prédictions)		4585 intersections non-nulles sur 10575 (43.4%) => 25.8 % rappel total
XGBoost (sortie : liste de 1000 prédictions)		8265 intersections non-nulles sur 10575 (78.2%) => 45.2 % rappel total

FIGURE 2.16 : Comparaison des scores de prédiction pour la deuxième approche.

3.4 Troisième approche

La troisième méthode est décrite en détail dans l'annexe A. Elle s'articule comme la construction d'un modèle de classement basé sur une méthode d'ensemble de type régression d'arbre de décision boosté (implémentation *LightGBM* du gradient boosting). Durant la phase d'apprentissage, l'ensemble des investisseurs est labellisé, suivant deux classes, ayant investi ou non, pour chacune des transactions du jeu d'entraînement du modèle. Compte-tenu de l'importance de la base de données et du déséquilibre important entre les classes, des sous-échantillonnages adaptés sont effectués pour une meilleure gestion des ressources de calculs, mais surtout pour rééquilibrer les classes. Lors de la prédiction, un score de classement, correspondant à la probabilité d'investir, est associé à chaque investisseur de la base de données, en fonction des caractéristiques de la transaction définies en entrée du modèle.

Arbre de boosting de gradient (implémentation LightGBM) – Dans cette approche on utilise LightGBM, qui est plus précis et plus performant en temps de calcul que XGBoost (*gradient boosting with scikit-learn*) mais il semble généralement moins utilisé car il est moins documenté [Ke et al., 2017]. Le développement se concentre sur les performances et l'évolutivité. En effet, LightGBM implémente un algorithme d'apprentissage d'arbre de décision basé sur un histogramme hautement optimisé, qui offre de grands avantages en termes d'efficacité et de consommation de mémoire. L'algorithme LightGBM utilise deux nouvelles techniques appelées *Gradient-Based One-Side Sampling (GOSS)* et *Exclusive Feature Bundling (EFB)* qui permettent à l'algorithme de s'exécuter plus rapidement tout en maintenant un haut niveau de précision. L'échantillonnage unilatéral basé sur le gradient (GOSS) est une méthode qui tire parti du fait qu'il n'y a pas de poids natif pour l'instance de données dans GBDT. Étant donné que les instances de données avec des gradients différents jouent des rôles différents dans le calcul du gain d'informations, les instances avec des gradients plus importants contribueront davantage au gain d'informations. Ainsi, afin de conserver la précision des informations, GOSS conserve les instances avec de grands gradients et supprime aléatoirement les instances avec de petits gradients.

Après un pré-traitement des données, le modèle est initialisé comme suit :

```
self.params = {
  'objective': 'binary',
  'boosting_type': 'gbdt',
  'metric': 'auc',
  'learning_rate': 0.1,
  'verbosity': -1
}

self.model = lgb.train(
  self.params,
  train_set,
  valid_sets=[valid_set],
  num_boost_round=10000,
  early_stopping_rounds=50,
  verbose_eval=50
)
```

On construit le *pipeline* suivant :

```
self.inputs = [
  'deal_type', 'series', 'vc_round', 'deal_size', 'percent_acquired',
  'company_id', 'primary_industry_sector', 'primary_industry_group', '
    primary_industry_code', 'company_hq_location', 'verticals'
]
self.pipeline = make_pipeline(
  Featurer(),
  DealCounter(),
  DealCounter(['deal_type']),
  DealCounter(['series']),
  DealCounter(['vc_round']),
```

```
    DealCounter(['company_id']),
    DealCounter(['primary_industry_sector']),
    DealCounter(['primary_industry_group']),
    DealCounter(['primary_industry_code']),
    DealCounter(['company_hq_country']),
    DealCounter(['company_hq_region']),
    DealCounter(['verticals']),
    DaysSinceLastDealCounter(),
    DaysSinceLastDealCounter(['deal_type']),
    DaysSinceLastDealCounter(['primary_industry_sector']),
    DaysSinceLastDealCounter(['primary_industry_group']),
    DaysSinceLastDealCounter(['primary_industry_code']),
    DaysSinceLastDealCounter(['company_hq_region']),
    SelectiveNumericalEncoder(self.categories),
    ColumnsSelector(columns_to_drop=[
    'deal_id', 'deal_date', 'invest',
    'company_id', 'company_name',
    'investor_id', 'investor_name', 'investor_hq_location'
    ])
  )
```

Comme évoqué précédemment, le détail des fonctionnalités de ce *pipeline* est présenté dans l'annexe A. On peut cependant noter l'importance de la prise en compte de la temporalité dans la création de caractéristiques qui intègrent l'occurrence des différentes caractéristiques au cours de l'année en cours (cf *DealCounter*), ainsi que du nombre de jours depuis la dernière transaction sur une caractéristique spécifique (cf *DaysSinceLastDealCounter*).

3.5 Comparaison des modèles

Les performances des trois méthodes ont été évaluées sur des données de tests. Le rappel (*recall*) est la métrique que nous avons retenue pour qualifier les performances de la méthode. Comme évoqué précédemment dans la section 1.5, il s'exprime comme :

$$\texttt{Rappel} = \frac{\texttt{Vrais Positifs}}{\texttt{Vrais Positifs} + \texttt{Faux Négatifs}} \tag{2.3}$$

On rappelle que les vrais positifs correspondent aux bons investisseurs qui se retrouvent dans la liste de 100 investisseurs suggérés par notre modèle. Les faux négatifs sont les investisseurs que le modèle n'a pas retenus alors qu'ils font en réalité partie du *deal*. Le rappel répond donc à la question : Quelle est la proportion d'investisseurs correctement identifiés par notre modèle.

Concernant l'évaluation des vrais positifs, les méthodes 2 et 3 sont plus performantes que la première puisqu'elles donnent des résultats de rappel moyen de 28% et 38% respectivement contre 17% pour la première. Contrairement à la littérature, nos résultats obtenus avec la première méthode utilisant un auto-encodeur sont décevants. La bonne performance du score de la troisième méthode est fortement liée à la création de nouvelles caractéristiques des investisseurs. La troisième l'approche présente une architecture intéressante avec la prise en compte de nouvelles caractéristiques faisant état de la chronologie comportementale des investisseurs, l'estimation des faux positifs s'avère supérieure à la deuxième méthode (10% de précision moyenne pour la 3ème contre 2% pour la 2ème). La troisième approche est donc la plus performante des trois méthodes, c'est la raison pour laquelle elle a fait l'objet d'un travail plus approfondi. L'objectif dans la suite du travail est de compléter la base de données avec des nouvelles caractéristiques pertinentes et d'augmenter la précision moyenne du système.

3.6 Analyse des résultats de la troisième méthode utilisant lgbm

Le tableau 2.1 illustre l'augmentation du score lorsqu'on prend un base de données élargie aux transactions de faibles montants, ainsi qu'en ajoutant de nouvelles *features* reflétant un historique des transactions et de l'activité des investisseurs. La moyenne sur toutes les prédictions de tous

les rappels moyens est globalement de 34-38% selon les cas. Notons que, malgré cette moyenne

Case 1	Case 2	Case 3	Case 4
Original Database + drop incomplete deals	Enlarged Database (including small deals) + drop incomplete deals	Enlarged Database (including small deals)	Enlarged Database (including small deals) + historical new features
Keeping 10487/18828 deals	Keeping 23048/61919 deals	Keeping 53412/61919 deals	Keeping 53412/61919 deals
25.5% mean Recall	27.8% mean Recall	33% mean recall	**37.5% mean recall**

TABLE 2.1 : Comparaison des scores pour la 3ème approche en fonction du nombre de *deals*, et de l'ajout de nouvelles caractéristiques prenant en compte l'historique des transactions.

d'environ 36% de bons investisseurs correctement identifiés, la variance est très élevée. On peut illustrer cette variance avec la figure 2.17 qui montre le rappel moyen en fonction du nombre d'investisseurs par transaction. En effet, il y a des *deals* sur lesquels le rappel est de 0% et d'autres 100%. À priori, on ne peut pas savoir si, dans la liste, il y a de bons investisseurs ou pas du tout. Il nous faudrait encore identifier les caractéristiques des *deals* qui offrent les moins bons résultats. Mais aucune caractéristique se dégage clairement. Le nombre d'investisseurs dans le *deal* ne semble pas être l'une d'entre elles.

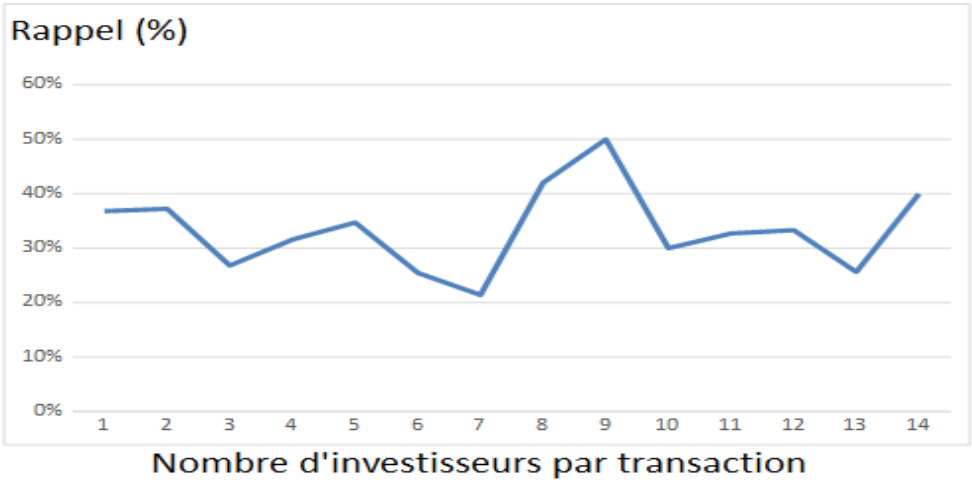

FIGURE 2.17 : Rappel moyen en fonction du nombre d'investisseurs par transaction

Interprétabilité – L'aspect "boîte noire" des modèles d'apprentissage apporte parfois une difficulté dans leur explicabilité. Ainsi, l'une des thématiques de recherche pour l'amélioration des moteurs de recherche et des système de recommandation concerne les méthodes d'explication des résultats, et d'interprétabilité. Pour comprendre les prédictions du modèle, l'interprétabilité consiste à comprendre et expliquer pourquoi certaines décisions ont été prises.

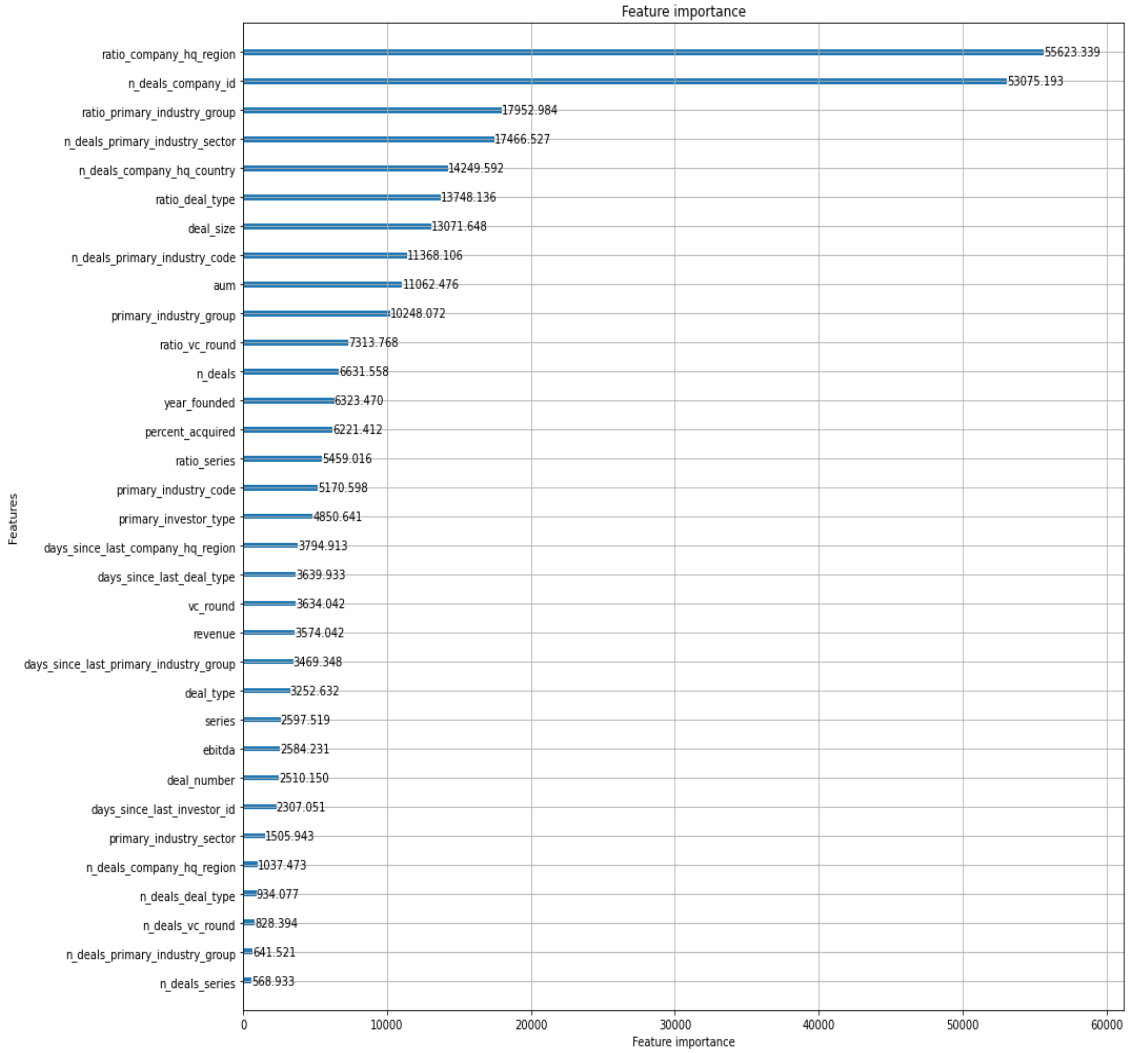

FIGURE 2.18 : Importance des caractéristiques.

La librairie lgbm possède une fonction intégrée qui donne directement l'importance des caractéristiques comme l'illustre la figure 2.18. On voit que les *features* les plus importantes sont le nombre de compagnies par région et le nombre de transactions par compagnie. Les *features* les moins importantes concernent les nombres de transactions par série, type de transaction ou secteur d'activité. On utilise aussi la librairie SHAP d'explication des caractéristiques (*i.e. SHAP, SHapley Additive exPlanations*)) telle que :

```
import shap
self.explainer = shap.TreeExplainer(self.model)
```

Elle quantifie la contribution que chaque *feature* apporte à la prédiction effectuée par le modèle. Pour ce faire, on calcule la contribution marginale de chaque *feature* apportée au résultat de prédiction du modèle afin de mettre en lumière l'importance des différents facteurs contribuant au choix des recommandation des investisseurs. En effet, mieux comprendre pourquoi une recommandation a été faite peut aider à en savoir plus sur le problème, les données et la raison pour laquelle le modèle peut échouer.

Une manière intuitive de comprendre la valeur de Shapley est l'illustration suivante : Les valeurs des caractéristiques entrent dans une pièce dans un ordre aléatoire. Toutes les valeurs de

caractéristiques de la salle participent au jeu, c'est à dire qu'elles contribuent à la prédiction. La valeur de Shapley d'une valeur de caractéristique est le changement moyen de la prédiction que la coalition déjà présente dans la salle reçoit lorsque la valeur de caractéristique les rejoint.

Les valeurs SHAP offrent deux grands avantages :

- **Interprétabilité globale** – Les valeurs SHAP peuvent montrer dans quelle mesure chaque prédicteur contribue, positivement ou négativement, à la variable cible. C'est comme le graphique d'importance des variables, mais il est capable de montrer la relation positive ou négative de chaque variable avec la cible.
- **Interprétabilité locale** – Chaque observation obtient son propre ensemble de valeurs SHAP. Cela augmente considérablement sa transparence. On peut expliquer pourquoi un cas reçoit sa prédiction et les contributions des prédicteurs. Les algorithmes traditionnels d'importance des variables ne montrent généralement les résultats que sur l'ensemble de la population, mais pas sur chaque cas individuel. L'interprétabilité locale nous permet d'identifier et de contraster les impacts des facteurs.

Lorsqu'on fait tourner le modèle, on extraie avec SHAP les cinq *features* les plus explicatives pour chaque investisseur présent dans la liste des 100 investisseurs. Par exemple, la figure 2.19 montre un classement des caractéristiques les plus fréquentes pour une transaction de l'entreprise Linkcy.

FEATURE	OCCURRENCE
Investor located in company state	96
Number of deals last year	90
Company state	76
Investor located in company region	58
Number of deals on country last year	34
Number of deals of deal type last year	28
Days since last deal on industry sector last year	24
Days since last deal on deal type last year	23
Days since last deal on region last year	21
Investor located in company city	19
Days since last deal last year	14
Investor AUM	6
Number of deals on verticals last year	6
Number of deals on region last year	2
Deal size (m$)	1
Investor age (in years)	1
Number of deals on industry sector last year	1

Figure 2.19 : Exemple de classement par occurrence des *features* les plus explicatives

Analyse des prédictions ne contenant aucun bon investisseur – Une analyse statistique est effectuée pour comprendre quelles sont les prédictions ne contenant aucun bon investisseur dans la liste ($Recall = 0$). Les figures 2.20 et 2.21 montrent les prédictions nulles en fonction du type de transaction. Les levées aux stade initial présentent le nombre le plus important de prédictions sans aucun bon investisseur (*Early Stage VC*). La figure 2.22 montre que cela est cohérent avec le nombre de prédictions nulles qui décroît progressivement au fur et a mesure de l'augmentation du nombre de tours de financement (de la série A à la série H). On peut donc comprendre que les caractéristiques intégrant l'historique des transactions pour une même société joue un rôle important dans la qualité des prédictions des investisseurs.

Analyse dealType

	recall = 0	recall = 0 (%)	Total	Total(%)	recall = 0/Total
Corporate	7	3 %	7	1 %	100 %
Early Stage VC	106	44 %	227	43 %	47 %
IPO	1	0 %	5	1 %	20 %
Later Stage VC	78	32 %	206	39 %	38 %
PE Growth/Expansion	32	13 %	53	10 %	60 %
PIPE	1	0 %	2	0 %	50 %
Secondary Transaction - Private	1	0 %	1	0 %	100 %
Seed Round	17	7 %	33	6 %	52 %
	243		534		

FIGURE 2.20 : Analyse par type de deal des prédictions sans aucun bon investisseur (rappel nul)

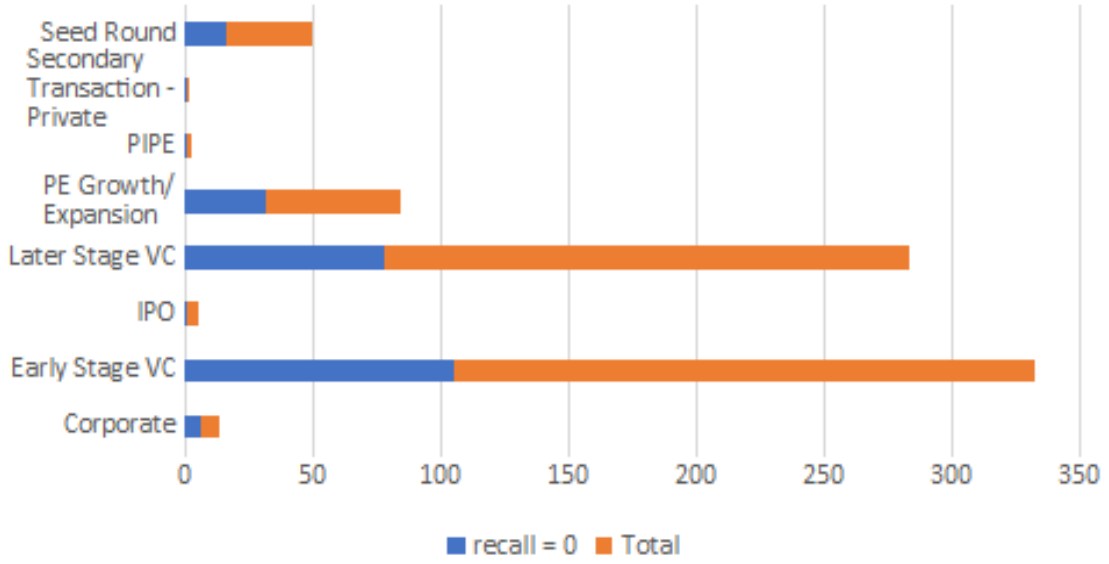

FIGURE 2.21 : Analyse par type de deal des prédictions sans aucun bon investisseur (rappel nul)

Analyse séries

	recall = 0	recall = 0 (%)	Total	Total (%)	recall = 0/Total
Series A	63	52 %	123	42 %	51 %
Series B	33	27 %	78	27 %	42 %
Series C	15	12 %	56	19 %	27 %
Series D	5	4 %	21	7 %	24 %
Series E	5	4 %	13	4 %	38 %
Series F	0	0 %	2	1 %	0 %
Series H	0	0 %	1	0 %	0 %
	121		294		

FIGURE 2.22 : Analyse par série des prédictions sans aucun bon investisseur (rappel nul)

Deal Size	recall = 0	recall = 0(%)	Total	Total(%)	recall = 0/Total
<$25m	73	58 %	127	51 %	57 %
$25-$50m	19	15 %	48	19 %	40 %
$50-$150m	21	17 %	47	19 %	45 %
$150-$500m	9	7 %	22	9 %	41 %
>$500m	3	2 %	6	2 %	50 %
	125		250		

FIGURE 2.23 : Analyse par taille de deal des prédictions sans aucun bon investisseur (rappel nul)

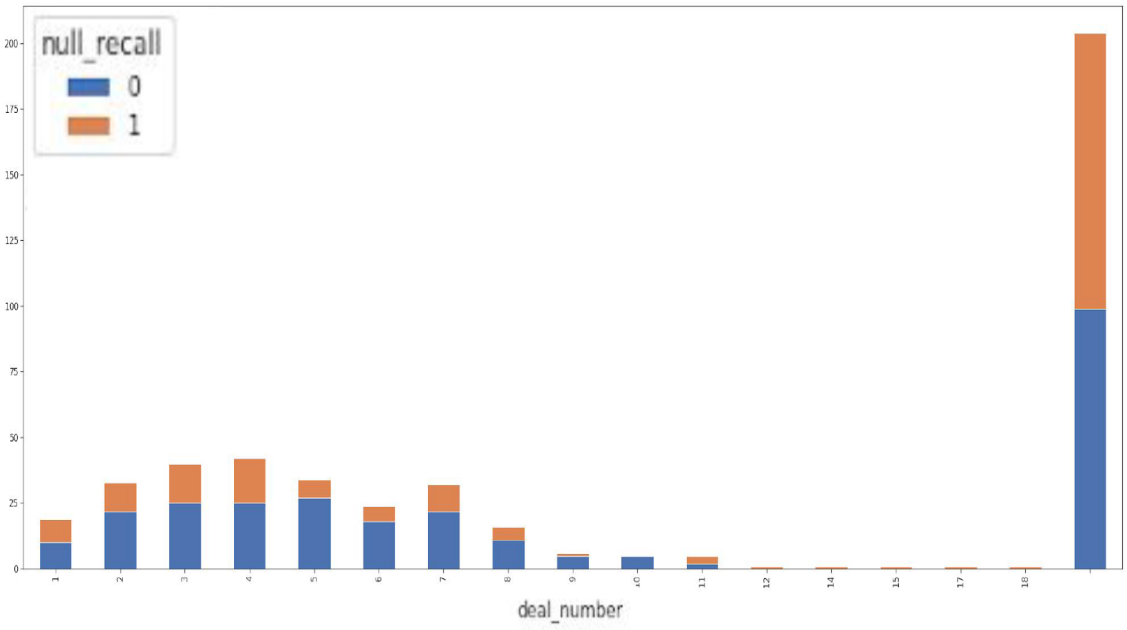

FIGURE 2.24 : Analyse des prédictions sans aucun bon investisseur (rappel nul) en fonction du nombre de deals précédents la transaction.

D'après l'analyse des *deal_size*, les transactions de taille <$25M semblent présenter de moins bons résultats (fig. 2.23). Parmi ces deals on trouve en particulier les *Seed Round* (en moyenne $15M et faible std). Les *Seed Round* représentent 5675 deals de notre base de 65204 deals, soit 8.7%. Les deals de taille <$25M représentent 21524 deals soient 33% des deals. À ce stade, le travail d'investigation des prédictions nulles peut être poursuivi en testant les résultats lorsqu'on enlève les transactions de moins de $25M et le *deal_number*. En effet, lorsque *deal_number* n'est pas renseigné, la figure 2.24 montre que la quantité de prédictions nulles est particulièrement conséquente. Enfin, de manière logique, la figure 2.25 montre que c'est lorsque le nombre d'investisseurs dans la transaction est particulièrement faible que l'on peut observer un grand nombre de prédictions sans aucun bon investisseur.

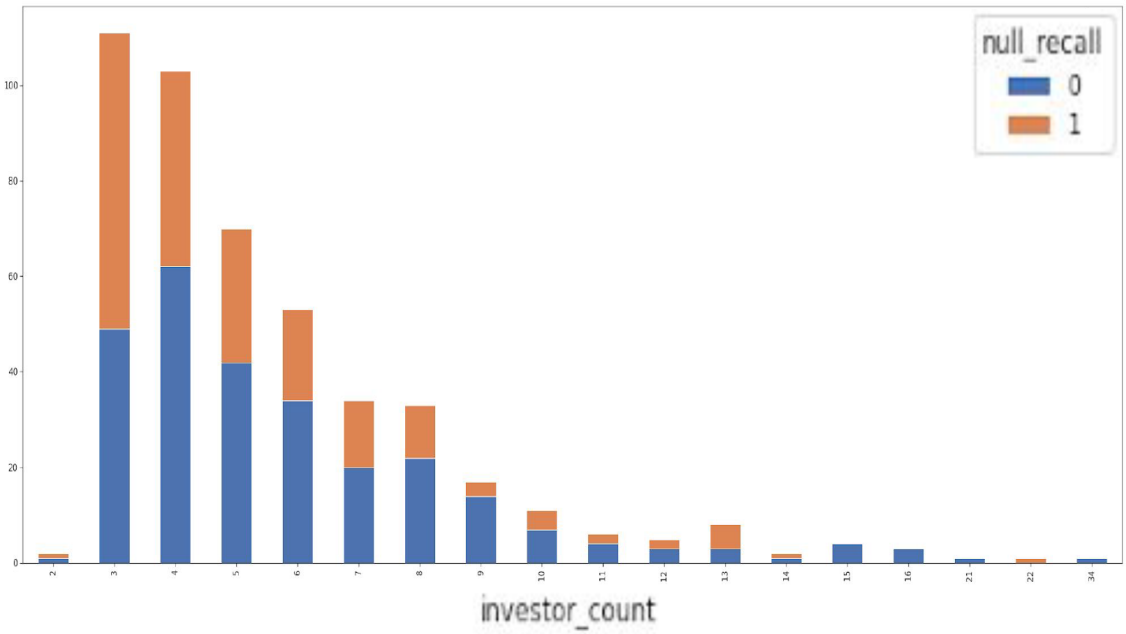

FIGURE 2.25 : Analyse en fonction du nombre d'investisseurs par deal des prédictions sans aucun bon investisseur (rappel nul)

Illustration d'un cas d'usage – L'étude des résultats des recommandations d'investisseurs a été réalisée, en particulier, sur différents cas d'usage. Parmi ceux-ci, des transactions ont été traitées concernant les sociétés suivantes : BLABLACAR, CHECKOUT, MIRAKL, BOXINE, PROXINVEST, YNSECT, ainsi qu'une société inconnue que l'on a nommée GREENFIELD.

FIGURE 2.26 : Scores du modèle et paramètres d'entrée pour la prédiction Mirakl.

Pour traiter ces cas, nous avons adopté la méthode suivante. On fait tourner le modèle avec plusieurs configurations de paramètres d'entrée. On fusionne ensuite les résultats issus de toutes les listes des meilleures recommandations pour chaque configuration en conservant, et en ordonnant, les probabilités correspondant aux scores de prédiction. À terme, afin d'éviter de lancer plusieurs simulations avec différentes configurations de transaction, on pourra améliorer la méthode en permettant au code d'intégrer dans la matrice de design les variations des paramètres d'entrée. Par exemple, si on teste le modèle pour le septième tour de financement de MIRAKL, d'un montant de \$300M, on prend en sortie les $2 \times 5 \times 3 = 30$ listes de 100 investisseurs prédits, associées aux combinaisons de paramètres d'entrée choisis dont la description est dans la figure 2.26. 190 investisseurs uniques sont trouvés et triés par probabilité.

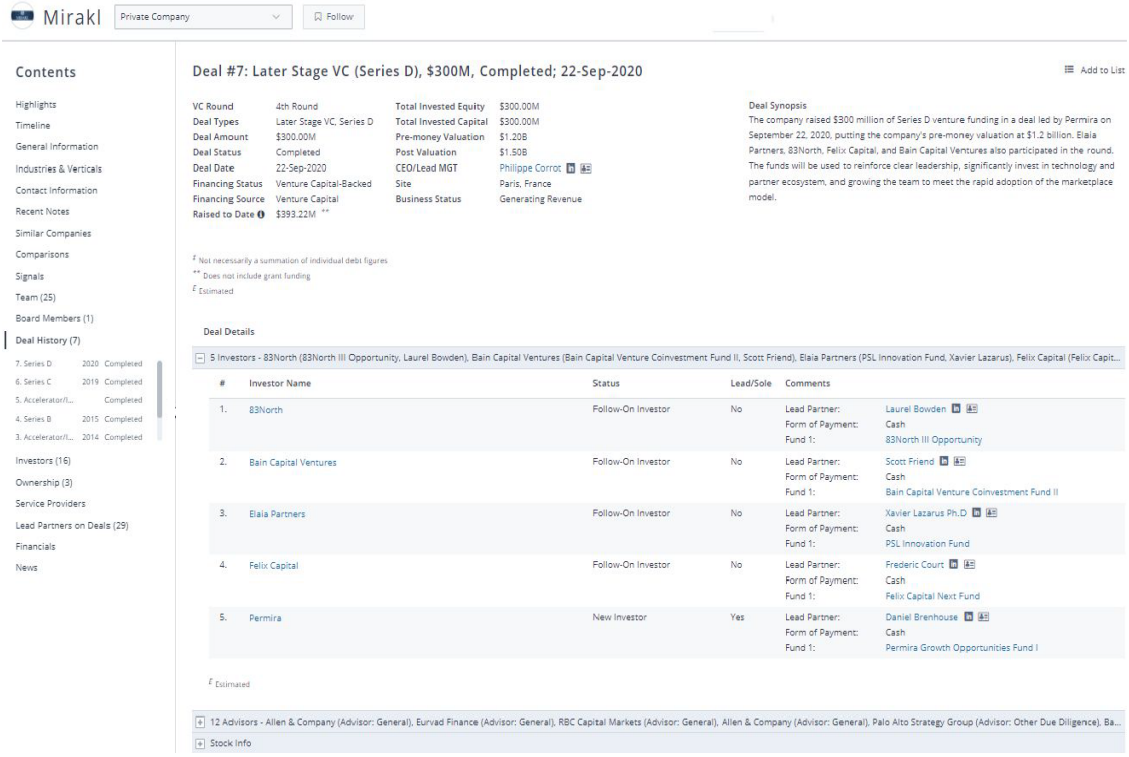

FIGURE 2.27 : Liste des 5 investisseurs participant au septième tour de financement de l'entreprise MIRAKL sur la plateforme pitchbook.

Les investisseurs réels de la transaction sont données par la figure 2.27. Parmi ces cinq investisseurs réels, on en trouve deux dans les prédictions (fig. 2.28). Le score de prédiction est bien conforme à nos attentes puisque le rappel moyen est donc de 40%. Notons tout de même que l'investisseur ELAIA PARTNERS présente une probabilité assez élevée puisqu'il se positionne à la $25^{\text{ème}}$ place dans le classement des recommandations. En outre, il apparait 18 fois au cours des 30 prédictions, il est donc bien récurrent. En revanche, l'investisseur PERMIRA n'apparait qu'une seule fois, et présente une position lointaine puisqu'il n'est qu'à la $118^{\text{ème}}$ place sur 190.

	investor_id	investor_name	primary_investor_type	aum	investor_hq_location	investor_hq_country	investor_hq_region	probability	occurrence
25	11170-63	Elaia Partners	Venture Capital	499.82	Paris, France	France	Europe	0.020360	18
118	11287-81	Permira	PE/Buyout	56407.71	London, United Kingdom	UK	Europe	0.009983	1

FIGURE 2.28 : Investisseurs corrects prédits participant au $7^{\text{ème}}$ tour de financement de MIRAKL.

4 Conclusions et perpectives

L'objectif à court terme dans la suite du travail est de compléter la base de données avec des nouvelles caractéristiques pertinentes et d'augmenter la performance du système. La suite des travaux consistera donc à tester des méthodes complémentaires pour améliorer les performances des prédictions ainsi que l'explicabilité des recommandations.

Modèle d'extraction – Le sous-échantillonage s'effectue sur l'ensemble des données et il semblerait pertinent, en amont de la prédiction, d'améliorer la récupération des données en améliorant le modèle d'extraction, notamment on pourrait faire un filtrage des investisseurs non pertinents pour la transaction pour optimiser ensuite l'étape de ré-équilibrage de la classe minoritaire.

Création de nouvelles *features* – La création de nouvelles *features* s'est révélé être une étape déterminante dans l'amélioration des prédictions. Ainsi, dans le prolongement de ce travail, on pourrait ajouter des caractéristiques qui caractérisent encore mieux les investisseurs. On pourrait par exemple ajouter une *feature* égale au rapport entre l'AUM (*i.e. Asset Under Management*) et le nombre d'investissements pour accéder au "ticket" moyen de chaque fond, ce qui revient à prendre en compte dans la transaction et la recommandation d'investisseurs associée, le montant de l'investissement moyen généralement aussi associée à la taille du fond d'investissement, ainsi qu'à la taille de son effectif.

Encodage des variables catégorielles – La manière d'effectuer l'encodage des variables catégorielles à champs multiples est également essentielle. Pour séparer toutes les valeurs des principales caractéristiques dans des colonnes individuelles, on pourrait tester l'encodage Multilabelbinarizer.

Les métriques – Le choix des métriques peut être amélioré et complété par l'utilisation du F1 score et du DGC comme cela été mentionné dans la section 1.5.

NLP – La figure 2.4 montre que la description des transactions (fig. 2.5) et des investisseurs (fig. 2.7) est généralement pratiquement toujours présente dans les données. Aussi, un traitement NLP qui impliquerait un "*word embedding*" pourrait permettre de rajouter de l'information utile dans les *features* utilisées dans le système de recommandation. Un *embedding* parfait conserverait des notions sémantiques de sorte par exemple que la relation : `king - queen = husband - wife` soit conservée [Rizkallah et al., 2021].

Séquentialité – Certains aspects temporels ont déjà été intégrés avec la création de nouvelles caractéristiques tenant compte d'un historique des transactions et des investisseurs. On a vu que cela améliore bien les prédictions. Pour encore mieux capter les tendances d'investissement, il serait intéressant de tester la prise en compte de la notion de séquentialité telle que cela est présenté dans certains articles [Zhao et al., 2020b ; Fang et al., 2020 ; Quadrana et al., 2018].

Chapitre 3

Classification des commentaires des investisseurs

À l'aide du NLP (*Natural Language Processing*) et dans le cadre de la collecte de données sur la plateforme de Praexo, l'objectif de ce projet est d'analyser, synthétiser et structurer automatiquement les commentaires des parties prenantes en un rapport efficace et intelligible, accessible à tous les utilisateurs impliqués dans la levée de capitaux.

1 Objectifs, vision et intégration du projet

Améliorer la communication de l'entreprise avec les investisseurs – Lorsqu'une entreprise se développe et traverse plusieurs étapes de financement, elle doit fournir de plus en plus d'éléments de transparence, incluant des rapports financiers approfondis, conformes à une régulation. En effet, le capital-risque a le potentiel de générer des rendements élevés, mais un investissement dans une entreprise en phase de démarrage est intrinsèquement plus risqué. Dans cet effort de transparence, le renforcement de la réglementation européenne, avec la mise en place de la MiFID 2 (*Markets in Financial Instruments Directive*), entrée en vigueur le 3 janvier 2018, offre une protection renforcée pour les investisseurs comparativement à la MiFID entrée en vigueur en 2007. Sa mission est de fixer les règles des institutions financières à propos de certains produits d'investissement ou de conseils. Ses objectifs sont les suivants : accroitre la transparence pour les clients et fluidifier l'exécution des transactions. Elle renforce l'obligation pour les institutions de disposer d'un suivi des audits, de la conformité et des risques en interne. Elle renforce la protection des investisseurs, la transparence et l'intégrité du marché. Les conseils en investissement sont désormais plus strictement encadrés.

Dynamiser l'intérêt de l'investisseur – Le top management doit obtenir un état instantané de son processus de levée de fonds, lui permettant d'anticiper les attentes et avis des investisseurs grâce à des messages et mises à jour facilement compréhensibles. Cela permet aux gestionnaires d'être plus percutants et plus précis, ce qui devrait stimuler l'attention des investisseurs.

Des décisions basées sur des analyses de données fiables – La plateforme de Praexo va devoir organiser un ensemble de données précieux grâce à son module d'analyse financière. Cela conduit à une meilleure découverte des prix et améliore le processus de prise de décision.

Principales caractéristiques – Le processus de retour d'information actuel qui canalise les opinions des investisseurs vers la direction via des intermédiaires est archaïque et rompu. Les ressources juniors des banques, des conseillers ou d'autres intermédiaires passent souvent d'innombrables heures à collecter manuellement les commentaires des investisseurs dans des fichiers

Excel et à préparer une analyse récapitulative dans des présentations PowerPoint. Le processus demande beaucoup de travail, est sujet aux erreurs et aux biais et souvent le produit final n'est pas entièrement compris par la direction de réception. En ouvrant un canal numérique direct entre les investisseurs et les entreprises, la plateforme de Praexo doit être en mesure de collecter, filtrer avec précision et rapporter des données plus rapidement et plus efficacement que jamais. Cela permet à toute personne impliquée dans le processus et ayant accès au tableau de bord de se concentrer sur des tâches à véritable valeur ajoutée. Le temps et les ressources peuvent être mieux alloués : au lieu de passer la majeure partie du temps dans une réunion à expliquer et comprendre les données de base, l'accent peut plutôt être mis sur le traitement des recommandations, la tactique et la résolution d'autres problèmes ou impasses.

La plateforme de Praexo doit proposer trois modules différents aux utilisateurs : un Dashboard, un module d'interface, et un modèle d'analyse financière.

Dashboard – Le tableau de bord va faciliter la communication et la collaboration entre la direction et ses conseillers. Toutes les parties auront accès à des informations en direct identiques concernant l'avancement et l'état de la transaction sur les marchés des capitaux, facilitant des interactions fluides et transparentes avec tous les participants à la transaction. Le tableau de bord sera disponible sur n'importe quel appareil : peu importe où se trouvent nos utilisateurs, ils resteront aux commandes. Le tableau de bord doit fournir :

- Des retours investisseurs synthétisés dans un format clair et pertinent avec des messages clés grâce à des techniques de data visualisation
- Un suivi en temps réel des métriques de valorisation et des KPI
- Une connaissance des investisseurs impliqués dans la transaction sur les marchés des capitaux
- Des mises à jour transparentes et en temps réel de la demande

Module d'interface – Le module d'interface peut être configuré pour répondre aux besoins spécifiques de tout client. Le front-end est une suite d'applications intégrées qui visent à rendre l'analyse des investissements et le processus de décision plus faciles et plus fluides pour les investisseurs.

Les investisseurs qui se connectent à la plateforme de Praexo auront accès à :

- Des informations de transaction enrichies en un seul endroit, y compris une installation de conférence Web pour les roadshows virtuels
- Une recherche indépendante en actions, y compris sur des sociétés privées grâce aux partenariats conclus par Praexo avec des sociétés de bourse de premier plan
- Un modèle d'analyse financière
- Un formulaire de commentaires standardisé avec outil voix-texte
- Un système de dialogue pour gérer efficacement les questions/réponses entre l'entreprise et les investisseurs
- Un système de réservation de commandes
- Une facilité de négociation pour les entreprises privées souhaitant initier des liquidités secondaires

Modèle d'analyse financière – La plateforme de Praexo va intégrer un modèle d'analyse financière unique qui permet aux investisseurs de contester le plan d'affaires de l'entreprise ou d'exécuter leur propre ensemble d'estimations individuelles avec la possibilité de comparer immédiatement les hypothèses avec les mesures d'exploitation des pairs et d'évaluer leur impact sur les notations.

Illustration de l'objectif du projet – L'analyse, la synthèse et la classification de commentaires des investisseurs doit s'intégrer au *dashboard*. Les histogrammes des figures 3.1, 3.2 et 3.3 illustrent un exemple de synthèse des forces, des faiblesses, et de la perception générale de la compagnie Elis, tel qu'on voudrait l'obtenir dans la plateforme de Praexo.

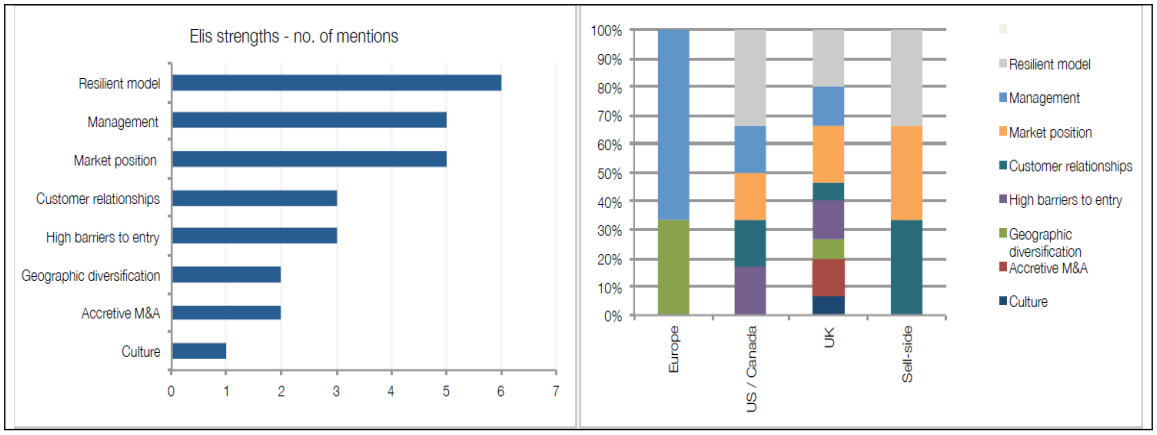

Figure 3.1 : Rapport des commentaires d'investisseurs concernant les forces de la compagnie Elis.

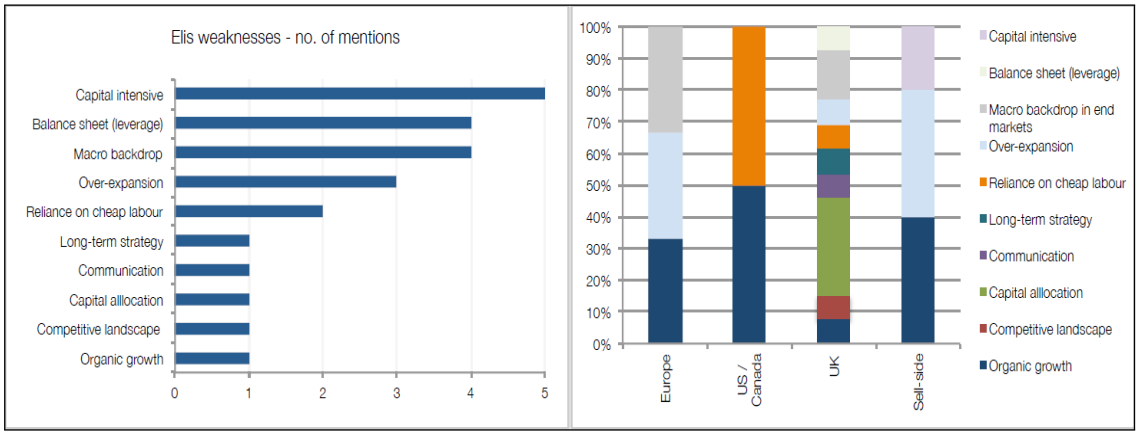

Figure 3.2 : Rapport des commentaires d'investisseurs concernant les faiblesses de la compagnie Elis.

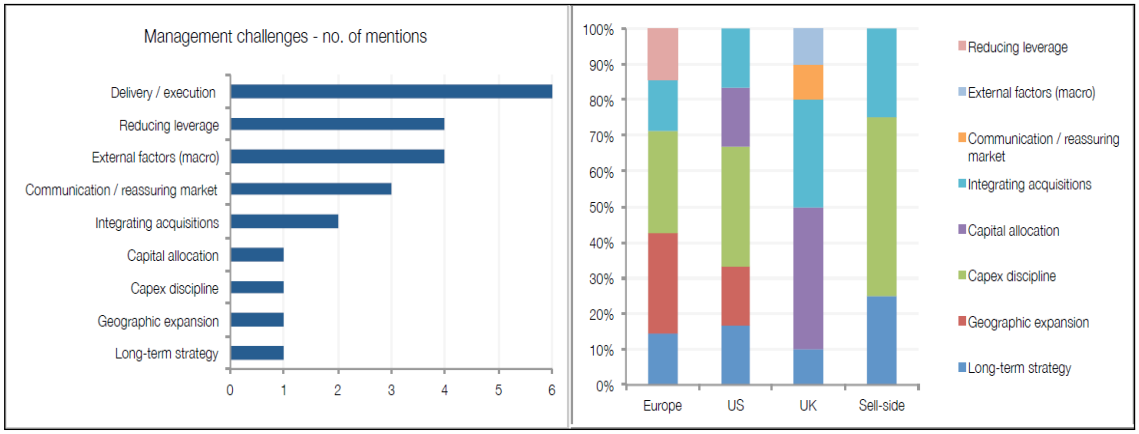

Figure 3.3 : Rapport des commentaires d'investisseurs concernant la perception générale de la compagnie Elis.

2 État de l'art

Adopter des ontologies dans les systèmes d'information a pour but de modéliser l'information au niveau sémantique [Brut and Sèdes, 2010]. La définition originale de l'ontologie au sens informatique a été fournie par T. R. Gruber en 1992, puis affinée par R. Studer [Studer et al., 1998] : "une ontologie est une spécification formelle et explicite d'une conceptualisation partagée d'un domaine de connaissances". L'acception actuelle considère l'ontologie comme une conceptualisation des connaissances d'un domaine particulier dans un format qui permet un traitement automatique, et qui modélise des entités, des attributs et des axiomes. Une ontologie est généralement composée de :

- Un vocabulaire dans lequel les termes désignent des concepts importants (classes des objets) du domaine ;
- Les relations entre les concepts comprennent habituellement des hiérarchies de classes ;
- Les propriétés des concepts ;
- Les limitations des valeurs ;
- Les spécifications des disjonctions ;
- La spécification des relations logiques entre les objets.

Une ontologie est donc définie comme suit : "Dans le contexte des sciences de l'informatique et de l'information, une ontologie définit un jeu de primitives représentatives avec lequel un domaine de connaissance ou un univers de discours peut être modélisé". Le terme "jeu de primitives" est la traduction la plus fidèle possible du monde réel à représenter. Une ontologie définit donc le vocabulaire partagé pour aboutir à une compréhension commune d'un domaine donné.

Nous devons donc construire une ontologie pour classifier les commentaires des investisseurs. Il existe des projets qui font l'effort de définir une base commune de terminologie et de concepts pour faciliter le traitement des données financières. Par exemple, l'onthologie des affaires de l'industrie financière (Financial Industry Business Ontology (FIBO)) tente de comprendre et homogénéiser la terminologie financière afin d'améliorer l'efficacité des décisions ou des rapports réglementaires et d'accélérer l'adoption des capacités analytiques avancées pour les services financiers.

La reconnaissance d'entités nommées (*NER i.e. Named Entity Recognition*) est une sous-tâche de l'activité d'extraction d'information dans des corpus documentaires et fait l'objet de nombreuses récentes recherches en NLP [Effland and Collins, 2021 ; Cao et al., 2018]. Les systèmes statistiques actuels utilisent une grande quantité de données pré-annotées pour apprendre les formes possibles des entités nommées. Il n'est plus nécessaire ici de rédiger de nombreuses règles à la main, mais d'étiqueter un corpus qui servira d'outil d'apprentissage. Ces systèmes sont donc eux aussi très coûteux en temps humain. Pour résoudre ce problème, récemment, des initiatives d'étiquetage automatique sont à l'étude [Jafari et al., 2020].

Dans les techniques traditionnelles de regroupement sémantique de documents, des mots ou phrases clés sont identifiés et traités pour évaluer la similitude du document, qui est une information cruciale pour effectuer le processus de regroupement. On peut aussi utiliser des entités nommées car elles peuvent compléter le clustering basé sur du texte brut pour rendre les résultats plus précis et significatifs. Le clustering basé sur entités nommées présente les avantages suivants.

- Les entités nommées peuvent être traitées comme des termes particuliers dans lesquels leurs significations et leurs relations sont précisément prédéfinies dans une ontologie. Ils donnent des résultats de clustering plus précis sur le plan sémantique pour certains besoins des utilisateurs.
- Dans certains domaines, comme les actualités publiées dans médias et matériels d'apprentissage, les entités nommées suggèrent des représentations ayant du sens pour les clusters générés parce qu'ils capturent des nœux sémantiques dans le contenu des documents.

Les étapes de base pour un document hiérarchique basé sur une entités nommées sont les suivants :

- Reconnaissance d'entités nommées : elle reconnaît les entités nommées dans les documents, paragraphes, phrases.
- Vectorisation des documents en fonction des entités nommées reconnues.
- Clustering hiérarchique multi-objectifs : il regroupe les vecteurs représentatifs en utilisant une mesure de similarité. Un cluster peut être divisé en plus petits clusters dans la hiérarchie.

Dans notre approche, on va utiliser un modèle de langage développé par Google en 2018 [Devlin et al., 2018]. Ce modèle, BERT (*i.e. Bidirectional Encoder Representations from Transformers*), est conçu pour pré-entraîner des représentations bidirectionnelles profondes à partir de texte non-étiqueté en conditionnant conjointement les contextes gauche et droit dans toutes les couches. En conséquence, le modèle BERT pré-entraîné peut être affiné avec une seule couche de sortie supplémentaire pour créer des modèles de pointe pour un large éventail de tâches. Le calcul de similarité entre phrases et catégories peut être optimisé en utilisant Sentence-BERT (SBert) [Reimers and Gurevych, 2019]. On peut donc résumer ainsi :

- création d'un corpus de validation à partir de notes de *brokers* ou de fonds d'investissement.
- développement de la classification avec Bert et la méthode de *zero-shot transfer learning* "zéro-shot"
- amorce d'un corpus étiqueté pour affiner le modèle pré-entrainé

3 Classification de commentaires d'investisseurs

3.1 Création d'un corpus

Pour constituer une base de données, on utilise des rapports de brokers et de fonds d'investissement au format `.pdf`. L'objectif est de catégoriser les *feedbacks* et uniformiser les informations financières des entreprises. Une librarie d'extraction de texte permet alors de constituer le corpus. Apache Tika est une bibliothèque utilisée pour la détection de type de document et l'extraction de contenu à partir de divers formats de fichiers. En utilisant cet outil, on peut développer un détecteur de type universel et un extracteur de contenu pour extraire à la fois du texte structuré et des méta-données de différents types de documents tels que des feuilles de calcul, des documents texte, des images, des PDF et même des formats d'entrée multimédia dans une certaine mesure. Tika-Python est une liaison Python aux services Apache Tika™ REST permettant à tika d'être appelé nativement en langage python.

3.2 Apprentissage par transfert

Lorsqu'on dispose de peu de données, l'apprentissage par transfert vise à transférer des connaissances d'une ou plusieurs tâches sources vers une ou plusieurs tâches cibles. Il peut être vu comme la capacité dun système à reconnaître et appliquer des connaissances et des compétences, apprises à partir de tâches antérieures, sur de nouvelles tâches ou domaines partageant des similitudes. L'idée de départ est donc de s'appuyer, quand c'est possible, sur un domaine adjacent, présentant une forte similarité avec le domaine à traiter, et pour lequel il est possible d'entraîner un modèle sur un grand jeu de données, que l'on appelle "données source". Bien sûr, un tel modèle donnera des résultats médiocres si on lui soumet directement un "jeu de données cible", mais il sera un bon point de départ. L'apprentissage par transfert va consister à utiliser des outils permettant dentraîner le plus efficacement possible ce modèle sur le jeu de données cible, malgré sa petite taille, en profitant au mieux de son entraînement initial sur le grand jeu de données source. L'apprentissage par transfert est une technique où les connaissances recueillies à partir d'une tâche ou d'un modèle sont utilisées dans une autre tâche de nature similaire. Le cas d'utilisation le plus évident de l'apprentissage par transfert est lorsque nous voulons modéliser la tâche A mais que

nous avons un petit ensemble de données, mais nous avons un grand ensemble de données pour une tâche similaire B. Nous construisons un modèle pour la tâche où nous avons une abondance de données, puis réutilisons les poids du modèle (en cas d'*embedding*) ou un sous-ensemble du modèle (généralement observé dans les problèmes de vision par ordinateur et plus récemment en NLP). On utilise donc la librarie transformers qui fournit des milliers de modèles pré-entraînés pour effectuer des tâches sur des textes tels que la classification, l'extraction d'informations, la réponse aux questions, la synthèse, la traduction, la génération de texte et plus encore dans plus de 100 langues [et al., 2020]. On utilise une méthode d'apprentissage par transfert "zéro-shot" à partir du notebook Zero Shot Pipeline sur colab [Xian et al., 2017].

À terme, comme on l'évoquera dans la section 3.3, ces modèles pré-entraînés sur un texte donné, peuvent être affinés sur nos propres ensembles de données [Liu et al., 2021]. La classification s'effectue en quelques lignes de codes :

```
from transformers import pipeline
classifier = pipeline("zero-shot-classification", device=0)
multiclass_flag = 'True'
data = classifier(row.sentence, candidate_labels, multi_class =
    multiclass_flag)
```

Description du pipeline – Dans ce pipeline, on utilise un modèle entraîné sur le corpus *Multi-Genre Natural Language Inference (MultiNLI)* qui est une collection participative de 433 000 paires de phrases annotées avec des informations d'implication textuelle. Le corpus est calqué sur le corpus *SNLI*, mais diffère en ce sens qu'il couvre une gamme de genres de textes parlés et écrits, et prend en charge une évaluation distinctive de généralisation inter-genre. La dernière couche du pipeline prédit l'un des trois labels : `contradiction, neutral et entailment` [Williams et al., 2018] pour qualifier et quantifier la nature contradictoire, neutre, ou liée de la paire formée par la phrase de commentaire (`row.sentence`) et la catégorie (`candidate.label`). Puisque nous avons une liste d'étiquettes candidates, chaque paire séquence / étiquette est alimentée par le modèle en tant que paire prémisse / hypothèse, et nous sortons les logits pour ces trois catégories pour chaque étiquette. Donc, pour une seule séquence, nous nous retrouvons avec une matrice de logits de forme `(num_candidate_labels, 3)`. On peut utiliser cette méthode pour une analyse de sentiment quantitative (positive/negative) avec une classification force/faiblesse.

Quand `multi_class=False`, on fait un softmax des entailment logits sur toutes les étiquettes candidates, c'est à dire

```
\texttt{logits[:,-1].softmax(dim=0)}.
```

Cela donne une probabilité pour chaque étiquette de sorte que leur somme soit égale à l'unité. La classification multi-classe est possible avec cette méthode (`multi_class=True`) pour traiter les sous-catégories.

Quand `multi_class=True`, un softmax est appliqué sur entailment vs contradiction pour chaque étiquette candidate indépendamment, c'est-à-dire

```
\texttt{logits[:,[0,-1]].softmax(dim=1)[:,-1]}.
```

Cela donne une probabilité pour chaque étiquette candidate entre 0 et 1, mais ils sont indépendants et ne totalisent pas 1. Quant au modèle d'hypothèse, il s'agit d'un modèle qui formate une étiquette candidate sous forme de séquence. Donc, s'il on passe d'une étiquette candidate de "politics" via le modèle et que le modèle d'hypothèse par défaut de "This example is about .", le modèle sera alimenté par "This example is about politics" en tant qu'hypothèse.

Une liste de 50 catégories cibles a été réalisée grâce aux équipes business afin de préparer la classification des commentaires des investisseurs. La figure 3.4 montre un extrait du fichier du fichier log du processing de la classification des commentaires d'investisseurs. La figure 3.5

montre les meilleurs scores de classement de paires commentaire-catégorie. Les classes les plus fréquemment attribuées sont représentées par la figure 3.6. Sur la figure 3.7, on représente les distributions de probabilité de chaque classe. Enfin la matrice de corrélation des classes (fig. 3.8) montre les relations entre les différentes catégories (fig. 3.9).

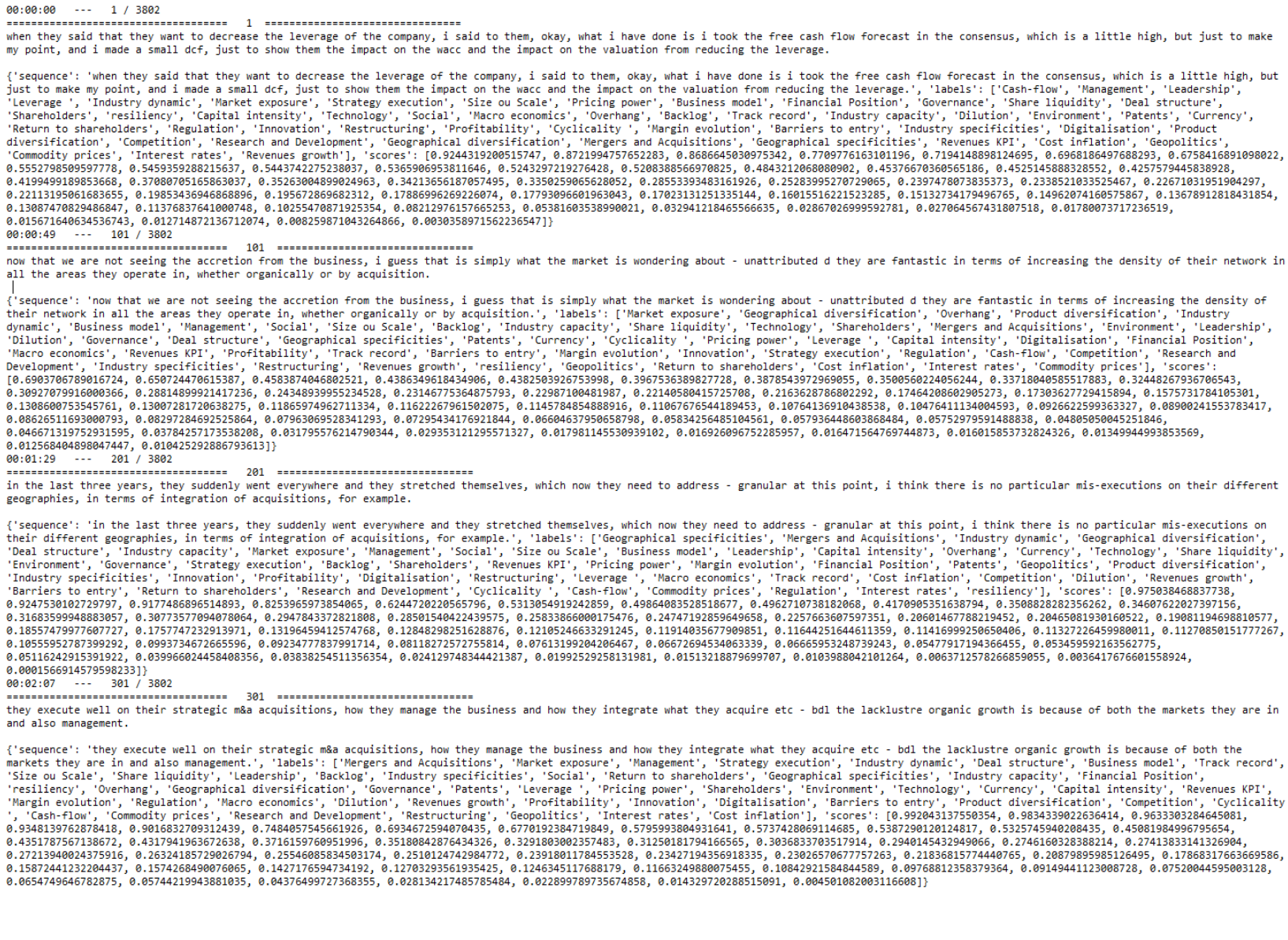

FIGURE 3.4 : Illustration du processing de commentaires d'investisseurs.

best_class	sentence	max_score
Cash-flow	these type of things, we can see it, but our first focus is in terms of cash flow.	84.0
Cost inflation	then there is the cost inflation, so they have to deal with some market topic.	84.0
resiliency	the second thing is probably the resilience, except probably the hospitality part of the business.	82.3
Governance	we rank governance quite high at 7.5, so 7 or 8 is fine for me.	81.0
Social	i am talking about the social background, which will impact the economy.	78.8
Leverage	leverage is one of these things where people think you should have more and more leverage and then they suddenly decide you should not have leverage and those two things can be a week apart.	78.0
Shareholders	in this area, we are shareholders, as you may know.	75.7
Capital intensity	i suppose a weakness you could point to on the strategy is that it is capital-intensive, although that is a barrier, so it can have a reasonable level of debt.	75.0
Financial Position	perception study - january 2019 financial position (cont.)	70.9
Revenues growth	not so good at organic revenue growth - dk partners i like them very much.	67.8
Commodity prices	the last negative is just in terms of raw materials; it does impact them, any movement in oil price or cotton price and stuff - franklin templeton we worry in general that elis is very focused on maximising its cost base.	66.2
Management	perception study - january 2019 management (cont.)	66.0
Mergers and Acquisitions	in order to get the top line going, you need to invest, and it is stuff that for the first few years just does not generate a lot of profitable margin and so he is just more focused on m&a.	64.8
Restructuring	it is just execution on the restructuring and not dropping the ball on the organic growth.	64.4
Strategy execution	on a scale of 1 to 10 how do you rate the execution of elis' strategy?	61.3
Business model	the first strength is probably the business model, but business model is quite a large word.	59.2
Interest rates	the point is more market perception and with interest rates rising and maybe market slowdown, there are many fears around high indebtedness.	58.9
Leadership	as i said at the beginning, it is a leader.	55.5
Market exposure	not on the management, but on the market, on this balance sheet structure, on the uk exposure, which is part of the business - lfde on the negatives, it is hard to know, berendsen might turn out to be brilliant or it might turn out to be a step too far.	54.2
Profitability	in terms of profitability, they have been doing quite well and everything is fine, according to me.	52.3
Geographical specificities	at this point, i think there is no particular mis-executions on their different geographies, in terms of integration of acquisitions, for example.	52.1
Track record	basically, if i summarise it, good execution in france, a good track record there of running the operations.	51.3
Pricing power	if they are unable to pass it through, there is a question around their pricing power and there is a question around the competitive dynamics, so that is another point.	51.3
Return to shareholders	forget dividends, i would do a buyback.	46.7
Industry dynamic	it is an industry in consolidation, then consolidation brings higher margins, higher returns.	46.4
Dilution	in terms of their strategy so far, the issue here is that they said we are going to consolidate markets and see margin expansion and i feel like that is being diluted a little bit.	46.1
Barriers to entry	you can see that with mainly distributors, like bunzl, but the barrier to entry is even higher, which is even more positive.	45.3
Competition	there is competition, but the berendsen brand is a strong brand, so there is no reason why it should not continue.	42.2
Industry specificities	there are no synergies, that is exactly what we like in this industry.	41.7
Cyclicality	they are going to face some cost pressures in the short-term, but we see that more cyclical.	41.1
Geographical diversification	regarding the geography, the business is more diversified now after the different acquisitions they made this past three years.	40.8
Macro economics	we will see how they are able to continue to grow in this different, maybe more complicated, macroeconomic context.	39.7
Margin evolution	it is a good mix between the organic sales growth and the margin evolution.	36.2
Size ou Scale	that goes then in pair with doing bolt-on acquisitions or a larger scale deal, but i do not think larger deals are on the horizon right now.	30.4
Share liquidity	that is the struggle the shares have at the moment.	28.4
Product diversification	regarding the business, they are quite balanced between workwear, linen, etc.	26.3
Currency	after that, it is also about the free cash flow generation and the roce.	24.1
Deal structure	keep doing small deals, rather than a very big structuring deal, leading to higher indebtedness and maybe capital increase.	23.3
Industry capacity	i think it is going to change now that, for example, they have hired a new ir.	21.1
Technology	seven, for their technical capacity.	18.3
Innovation	for example, for me, indusal was a material acquisition.	17.7
Revenues KPI	there is only a revenue number, so i do not really have much to say about it.	15.6
Environment	you are in a different environment.	13.6
Overhang	uk, that is a question mark.	11.9
Geopolitics	the prospects of a more challenging macro economic / geopolitical environment does not help.	10.9
Backlog	in uk, i hope they could restore the growth base in the future.	10.1
Regulation	i cannot express these kind of views for compliance reasons.	6.7

FIGURE 3.5 : Illustration de commentaires d'investisseurs avec leur résultat de classification de catégorie.

```
output.best_class.value_counts()
```

```
Industry dynamic                621
resiliency                      467
Market exposure                 323
Management                      289
Social                          269
Mergers and Acquisitions        213
Capital intensity               106
Cash-flow                       104
Pricing power                    90
Geographical specificities       89
Geographical diversification     79
Leverage                         79
Shareholders                     73
Business model                   73
Track record                     71
Size ou Scale                    65
Financial Position               64
Leadership                       59
Overhang                         59
Return to shareholders           56
Environment                      51
Profitability                    50
Interest rates                   49
Share liquidity                  37
Barriers to entry                36
Strategy execution               34
Deal structure                   33
Industry specificities           31
Technology                       29
Margin evolution                 29
Cost inflation                   25
Governance                       22
Competition                      21
Macro economics                  15
Dilution                         14
Revenues growth                  12
Product diversification          11
Cyclicality                      10
Innovation                        8
Restructuring                     7
Currency                          6
Backlog                           6
Geopolitics                       5
Industry capacity                 5
Commodity prices                  3
Revenues KPI                      3
Regulation                        1
Name: best_class, dtype: int64
```

FIGURE 3.6 : Occurrence des classes de commentaires d'investisseurs.

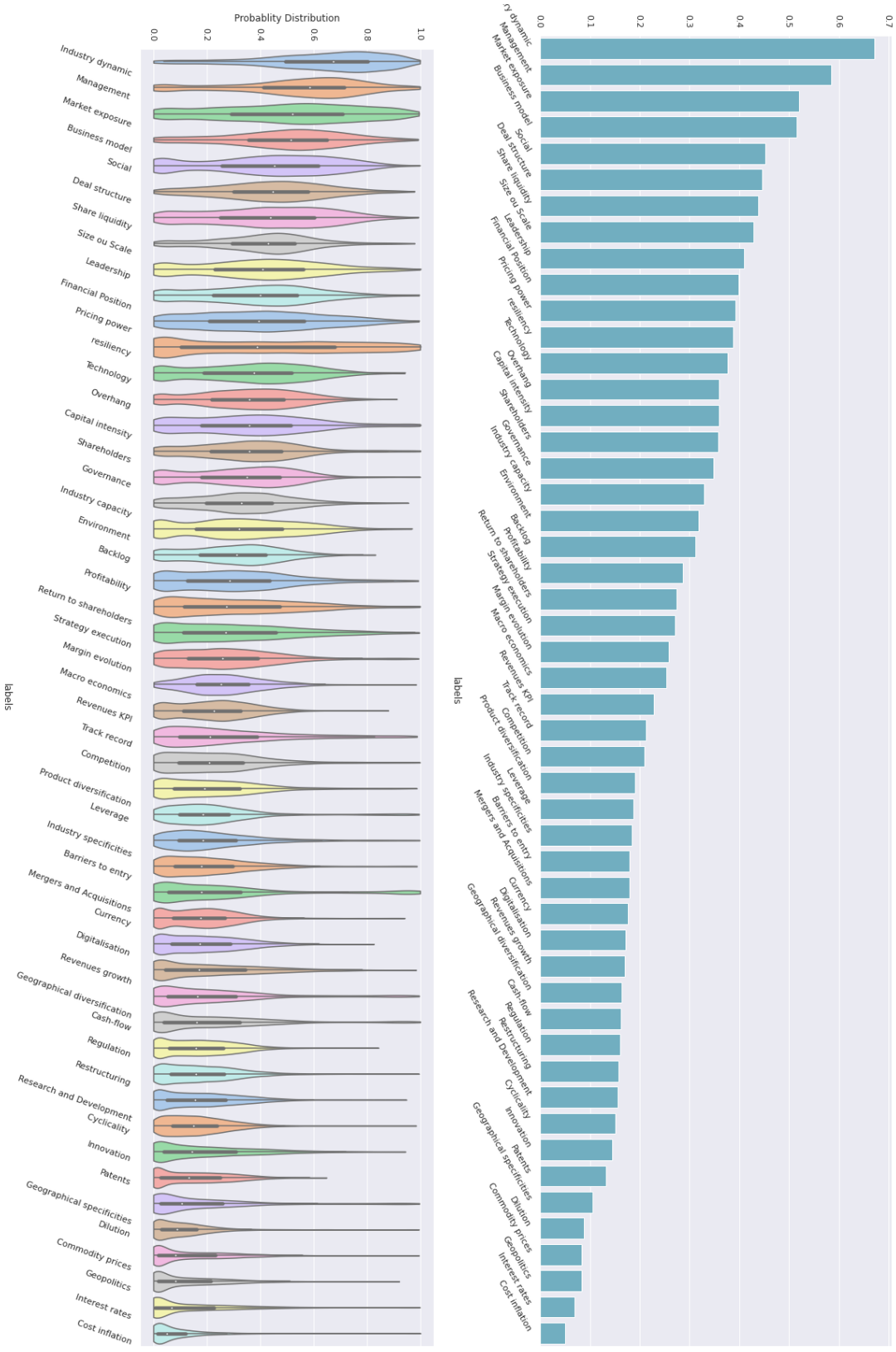

FIGURE 3.7 : Distribution de probabilités et médiane de chaque classe.

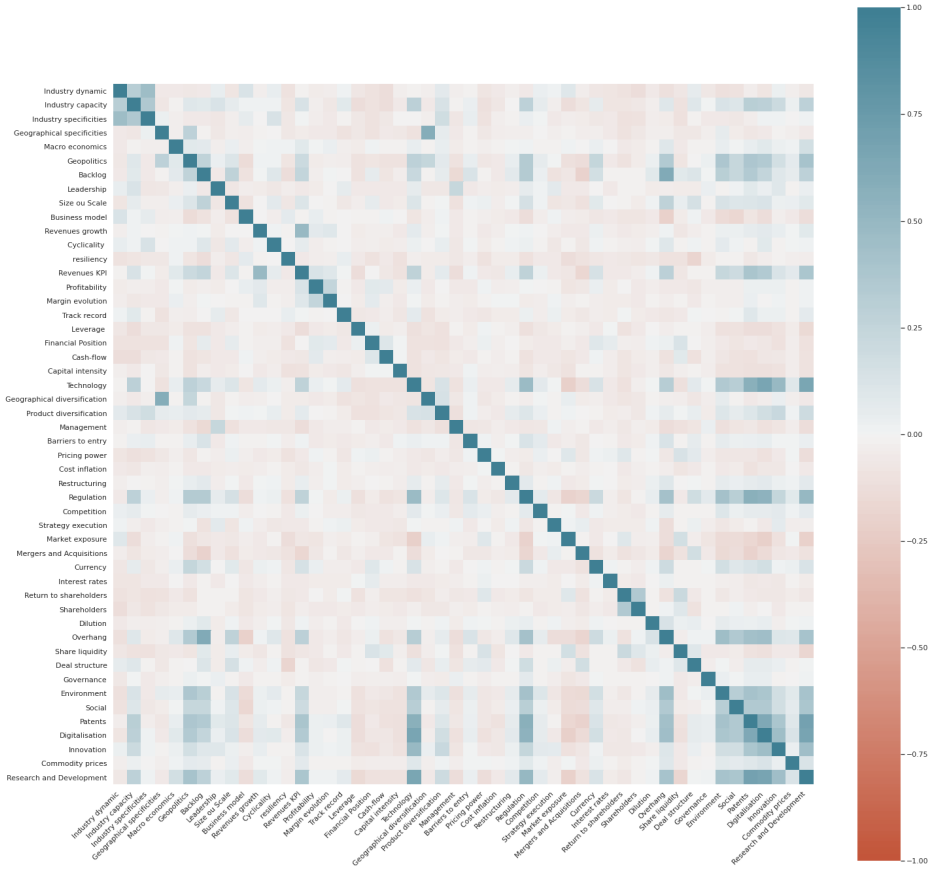

FIGURE 3.8 : Matrice de corrélation des catégories.

Most correlated class pairs

Revenues KPI	Revenues growth	0.7830289691342903
Geographical diversification	Geographical specificities	0.774720089536774
Governance	Regulation	0.7679672082286035
Management	Leadership	0.7559184187951488
Technology	Digitalisation	0.7500495209057306
Regulation	Patents	0.7324591410197278
Digitalisation	Research and Development	0.7274521592336403
Social	Environment	0.7270856450559333
Overhang	Backlog	0.6914424244046473
Technology	Research and Development	0.6810353669501931
Profitability	Margin evolution	0.6809309219202768
Digitalisation	Regulation	0.6793363992138026
Digitalisation	Innovation	0.677047769743367
Technology	Innovation	0.6754948375998939
Business model	Deal structure	0.6741314981820912
Digitalisation	Patents	0.6681078647858468
Social	Governance	0.6649479016762536
Regulation	Research and Development	0.662362854865845
Product diversification	Geographical diversification	0.6616076534196779
Environment	Regulation	0.656419002264126
Governance	Environment	0.6554063184260114
Governance	Patents	0.6474567444836278
Product diversification	Innovation	0.6446970982462443
Market exposure	Share liquidity	0.6411087437404279
Backlog	Patents	0.6340535124660625
Social	Technology	0.6329797327345499
Research and Development	Patents	0.6325252850509565
Currency	Patents	0.6307690971164484
Size ou Scale	Backlog	0.6296438823205713
Share liquidity	Pricing power	0.624483753585978
Management	Governance	0.6241608193820373
Leadership	Governance	0.6238142392115258
Technology	Patents	0.6213454981654551
Technology	Industry capacity	0.6211870587100459
Leadership	Strategy execution	0.6140483914155709
Shareholders	Return to shareholders	0.6127347089986849
Social	Regulation	0.6123471434749413
Deal structure	Pricing power	0.6111098796397327
Market exposure	Pricing power	0.6053052136195095
Technology	Governance	0.6037729820632515
Governance	Digitalisation	0.603642750694168
Industry dynamic	Deal structure	0.6035274199275028
Research and Development	Innovation	0.6022974841646918
Environment	Patents	0.6011331134259494
Backlog	Currency	0.5978904603136443
Environment	Research and Development	0.5973200707273835
Technology	Regulation	0.5950360504504445
Leadership	Industry capacity	0.5934353412901684
Share liquidity	Financial Position	0.5932898460903281
Currency	Regulation	0.5923394714440636
Governance	Research and Development	0.5909924395433013

FIGURE 3.9 : Catégories les plus corrélées.

3.3 Classification hiérarchique multi-classe

La classification hiérarchique multi-classe est basée sur l'utilisation de catégories hiérarchisées et de l'utilisation d'une structure de dendogramme [del Moral et al., 2021 ; Wang et al., 2021 ; Kowsari et al., 2017]. La méthode s'inspire de la méthode classifyText de `Google NLP` en définissant une liste de catégories hiérarchisée en trois niveaux du plus général au plus spécifique afin de permettre une certaine granularité dans la classification. Idéalement, on souhaite les catégories du premier niveau très générales, celles du deuxième, plus précises, et celles du troisième, si elles sont définies, aussi spécifiques que possible et mutuellement exclusives.

L'objectif du modèle de classification est de classifier un commentaire dans le niveau de plus forte granularité en premier lieu ($3^{\text{ème}}$ niveau), ensuite dans le $2^{\text{ème}}$ niveau s'il n'y a pas de correspondance avec le $3^{\text{ème}}$ niveau, et enfin dans le 1^{er} niveau s'il n'y a pas de correspondance dans le $2^{\text{ème}}$ niveau. Par exemple, si `/Financials` et `/Financials/Shareholder` s'applique tous les deux à un commentaire, alors seulement la catégorie `/Financials/Shareholder` sera retournée puisque c'est un résultat plus spécifique. Cette méthode permettra de représenter les catégories de chacun des niveaux dans l'interface utilisateur.

Dans la suite, il est possible d'ajouter plus d'éléments aux catégories de troisième niveau mais les catégories de niveaux 1 et 2 doivent être moins flexibles. Idéalement, afin de faciliter la classification multi-classes, il est souhaitable d'avoir une dizaine de catégories de premier niveau tout au plus. Notons également que la classification est globalement plus efficace lorsque les catégories ne comportent qu'un seul mot plutôt que plusieurs.

Labélisation des catégories hiérarchisées pour la classification – Les catégories les plus générales, pour le premier niveau, sont :

- Environment
- Industry specifics
- Market
- Business
- Assets
- Financials
- Valuation
- Management
- Strategy
- ESG criteria
- Deal

Les catégories multi-classes retenues sont décrites en détail dans l'annexe 1. L'ensemble des catégories est reportée jusqu'à quatre niveaux de sous-catégories.

Entraînement d'un modèle spécifique La création d'un corpus étiqueté spécifique aux commentaires des investisseurs a été amorcée dans le but de permettre d'entrainer un modèle mieux adapté à notre utilisation. Ainsi, afin d'augmenter la performance de la classification des commentaires des investisseurs, l'entraînement d'un modèle à partir de données spécifiques labellisées va pouvoir être effectué. Basée sur les librairies spaCy, Prodigy est une interface d'annotation des données qui intègre l'apprentissage actif d'un modèle. Cela permet d'accélérer et d'améliorer les performances du processus de classification. Le principe est décrit par les étapes suivantes :

- création d'exemples de phrases bien classifiées
- création d'une collection d'exemples labellisés à l'aide des exemples de phrases bien classifiées
- entraînement d'un modèle temporaire
- étiquetage d'exemples supplémentaires en corrigeant le modèle d'entraînement
- entraînement d'un meilleur modèle en améliorant la précision
- application du modèle sur tous les commentaires.

Si nous créons un bon ensemble de données de validation étiquetées, nous pourrons affiner un modèle, pour obtenir de meilleures performances à un coût de calcul inférieur.

Chapitre 4

Conclusions et perspectives

Après sa première levée de fonds de 1.6M€ en mai 2020, ce travail s'inscrit dans l'amorçage de l'activité de la jeune fintech Praexo, avec le développement de prototypes, de produits minimum viables, ainsi que de premières solutions techniques commercialisables. Au cours de ce travail, deux projets distincts ont été abordés. Le premier concerne la mise au point d'un système de recommandation d'investisseurs pour les levées de fonds d'entreprises. Le deuxième projet a seulement été initié et concerne une analyse, une synthèse et une classification des commentaires d'investisseurs.

Si les modèles de machine learning sont au cœur des algorithmes de recommandation, ils n'en sont pas les uniques constituants. À l'algorithme, il faut plutôt substituer une vision d'un ensemble d'algorithmes travaillant de concert, mêlant modèles avancés de machine learning, règles métiers et traitement massif des données d'entrée. Chaque étape est le fruit de compromis techniques, scientifiques mais avant tout humains : comment interpréter ces données ? Comment choisir les modèles ? Comment les mettre en œuvre pour qu'ils répondent aux contraintes techniques d'une plateforme de services de financement d'entreprise ? À chaque étape, ce sont les décisions des différents acteurs à l'œuvre (product manager, data scientists, développeurs, designer, banquiers d'affaires...) qui vont produire du sens et transformer un objet technique fait d'informatique et de mathématiques en un outil au service des entreprises et des investisseurs. Dès lors, il s'agit de coordonner le travail entre ces acteurs et permettre de réaliser des algorithmes, d'itérer et coupler les dernières avancées de l'état de l'art en machine learning avec des contraintes de mise en production rapide. Cette organisation doit relever des défis majeurs : adapter le cadre agile à un rythme et une incertitude liée aux problématiques de data science, tirer meilleur parti des spécialisations de chacun, tout en œuvrant à un but commun, aligner les cultures variées des différents métiers d'une entreprise à la prise de décision assistée par la donnée. Mais au final, quelque soient les choix discutés et proposés, une seule personne aura le dernier mot : l'utilisateur.

1 Système de recommandation d'investisseurs

Conclusion – Le scraping des données a été réalisé sur la plateforme pitchbook.com qui contient des données de transactions pour le financement des entreprises à l'échelle globale. Pour réaliser le système de recommandation des investisseurs, trois approches totalement différentes ont été développées : la première basée sur un auto-encodeur, la deuxième sur une SVD, et la troisième sur un arbre de décision à gradient boosté. Les meilleurs résultats ont été obtenus avec la troisième approche, en particulier grâce à la création de nouvelles caractéristiques qui intègrent des statistiques telles que des rapports de *features*, des comptages et des séquences témoignant de l'historique des transactions.

Les résultats de cette première mouture de système de recommandation d'investisseurs ont été

particulièrement bien accueillis par une communauté de professionnels qui ont valeur de "valideurs", et de premiers clients tels que BNP Paribas. Plusieurs cas d'usage concernant différentes transactions de sociétés telles que Ynsect, Proxinvest, Boxine, Mirakl, Checkout, ou Blablacar ont été réalisés afin de permettre un retour d'information primordial pour réaliser et valider un produit fini plus mature.

Perspectives – Parmi les nombreuses problématiques que nous avons pu identifier, auxquelles il s'agira d'apporter des solutions à la suite de ce travail, on peut citer les points suivants.

- L'amélioration du scraping de la base de données a été amorcée, autant dans la quantité de données scrapées, que dans la fréquence du scraping. Cela devrait permettre de faire une analyse statistique détaillée puis de consolider les résultats. Notamment, le nettoyage et l'homogénéisation des données doit être bien pris en considération. Il y a dans la base de données actuelle certaines erreurs difficilement décelables automatiquement, par exemple, un investisseur allemand qui n'est pas dans le bon secteur. À l'issue de cette première étape de prototypage, une solution pérenne d'extraction des données doit être trouvée, à terme, pour générer une base de données robuste dans le cadre d'une solution commercialisable. En particulier, la mise à l'échelle et l'industrialisation du code suppose de réaliser une mise à jour régulière de la base de données, idéalement chaque jour. Enfin, un effort doit être fait pour fusionner les différentes sources de données qu'elles soient internes à Praexo, ou bien externes, avec l'utilisation de multiples bases de données.
- Une attention particulière doit être portée au modèle d'extraction des données que ce soit dans l'étape de pré-traitement, ou bien en amont de la prédiction. En effet, le filtrage des données en sous-ensembles peut permettre, par exemple, de faciliter le ré-équilibrage des classes dans l'étape de sous-échantillonnage.
- Bien que la précision des solutions proposées soit très satisfaisante, la principale problématique concerne la présence de nombreuses sorties non-pertinentes. Les algorithmes doivent minimiser l'occurrence de faux-positifs dans les résultats de recommandation. Suggérer des investisseurs non-pertinents se révèle être un problème business crucial pour notre application. Il n'a pas suffisamment été considéré en amont de ce projet. Pour remédier à cela, on peut soit intégrer une métrique adaptée au modèle actuel, soit améliorer l'*embedding*. Il est cependant indispensable d'effectuer une bonne caractérisation des investisseurs non-pertinents grâce à l'expertise des équipes business.
- Comme évoqué dans la section 4, l'utilisation de combinaisons de métriques additionnelles peut contribuer grandement à l'amélioration des résultats.
- Le travail sur l'encodage des variables catégorielles et l'*embedding* est également essentiel et doit faire l'objet d'une analyse en profondeur. À ce titre un traitement NLP de *word embedding* peut être envisagé afin de mieux traiter l'information contenue dans les variables catégorielles.
- La création de nouvelles variables explicatives, prenant en compte, en particulier, l'historique ou des statistiques sur d'autres *features*, s'est avérée bénéfique. On a pu constater, par exemple, que l'AUM (*Asset Under Management*) n'était pas suffisant pour caractériser le fond d'investissement et le montant moyen de sa participation aux transactions. L'intégration de l'effectif du fond, ou la création de nouvelles *features*, peuvent renforcer les propriétés de son profil afin de ne pas le suggérer dans des transactions inadéquates.
- Une autre idée serait de mieux prendre en considération l'aspect séquentiel des investissements de chaque investisseur et entreprise au fur et à mesure que la base de données s'étoffe.
- L'influence contextuelle relative aux critères géographiques doit être mieux maîtrisée. En effet, dans le produit fini, on doit pouvoir demander au système une flexibilité dans le paramétrage afin d'obtenir des recommandations plus ou moins localisées dans certaines régions, en fonction du choix de l'utilisateur.

- Le travail sur l'explicabilité des recommandations doit être poursuivi car il est une clé essentielle pour parvenir à interpréter et améliorer les recommandations.

2 Classification des commentaires d'investisseurs

Conclusion – Une première étape a été effectuée dans l'analyse et la synthèse des commentaires d'investisseurs. Afin d'alimenter les tableaux de bord de la plateforme de Praexo, synthétisant l'information des différentes entreprises et leur état de financement, on s'est confronté à la problématique de la classification automatique de texte. Après avoir créé un corpus propre, non-labellisé, à partir de commentaires d'investisseurs, notes de brokers, analyses de fonds d'investissements, les phrases ont été extraites et étiquetées avec un modèle pré-entrainé et une méthode d'apprentissage *zero-shot*. La génération d'entités automatiques n'ayant pas donné de résultats concluants, nous avons utilisé des noms de catégories déterminés par l'expertise des équipes business car elle correspond bien aux préoccupations ciblées. Cette méthode flexible permet d'ores et déjà d'effectuer une analyse de sentiments en commentaires favorables ou défavorables, correspondant à la perception de l'investisseur par rapport aux forces et faiblesses de l'entreprise. Elle délivre un score de similarité entre toutes les paires phrase de commentaire–catégorie cible. Elle intègre également la possibilité de modifier ou d'ajouter des catégories au fil du temps, et d'avoir un commentaire classifié dans plusieurs catégories.

Perspectives – Un travail important reste à réaliser. L'objectif est de parvenir à une classification hiérarchique intégrant la granularité des sous-catégories mutuellement exclusives. Parmi les travaux en cours, un étiquetage manuel d'une partie du corpus va permettre de se servir de ces données d'apprentissage pour affiner le modèle pré-entrainé et quantifier la performance des différentes approches. À ce titre, il sera important de considérer le choix d'une métrique adaptée aux objectifs désirés. Enfin, une liste exhaustive de catégories à quatre niveaux hiérarchiques a été établie et servira de base pour les travaux futurs.

Bibliographie

D. Dhillon, M. Granzer, T. Jennison, and U. Zier. Investment banking in a machine age. *Boston Consulting Group*, sept. 2020.

IRmagazine. Investor targeting. Technical report, IRmagazine.com, 2019. URL https://content.irmagazine.com/story/investor-targeting-research-report-sample/.

Zeinab Shahbazi, Debapriya Hazra, Sejoon Park, and Yung Cheol Byun. Toward improving the prediction accuracy of product recommendation system using extreme gradient boosting and encoding approaches. *Symmetry*, 12(9), September 2020. doi : 10.3390/SYM12091566. URL https://pdfs.semanticscholar.org/a4c0/339ecf404574918d38305630f44e4f89f73a.pdf?_ga=2.146802357.382257103.1631184058-1600487016.1629822383.

Idir Benouaret. *Un système de recommandation contextuel et composite pour la visite personnalisée de sites culturels. (A contextual and composite recommender system for the personalization of cultural sites visit)*. PhD thesis, University of Technology of Compiègne, France, 2017. URL https://tel.archives-ouvertes.fr/tel-01767997.

Sihem Amer-Yahia and Idir Benouaret. A Comparative Evaluation of Top-N Recommendation Algorithms : Case Study with Total Customers. In *IEEE Big Data*, Now Taking Place Virtually, France, 2020. URL https://hal.archives-ouvertes.fr/hal-03002602.

Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. *Recommender Systems : An Introduction*. Cambridge University Press, 2010. doi : 10.1017/CBO9780511763113.

Zhi-Peng Zhang, Yasuo Kudo, Tetsuya Murai, and Yong-Gong Ren. Enhancing recommendation accuracy of item-based collaborative filtering via item-variance weighting. *Applied Sciences*, 9(9), 2019. ISSN 2076-3417. doi : 10.3390/app9091928. URL https://www.mdpi.com/2076-3417/9/9/1928.

J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez. Recommender systems survey. *Knowledge-Based Systems*, 46 :109–132, 2013. ISSN 0950-7051. doi : https://doi.org/10.1016/j.knosys.2013.03.012. URL https://www.sciencedirect.com/science/article/pii/S0950705113001044.

Julien Delporte. *Factorisation matricielle, application à la recommandation personnalisée de préférences*. PhD thesis, INSA, 2014. URL http://www.theses.fr/2014ISAM0002/document. Thèse de doctorat dirigée par Canu, Stéphane Informatique Rouen, INSA 2014.

Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Kaiyuan Li, Yushuo Chen, Yujie Lu, Hui Wang, Changxin Tian, Xingyu Pan, Yingqian Min, Zhichao Feng, Xinyan Fan, Xu Chen, Pengfei Wang, Wendi Ji, Yaliang Li, Xiaoling Wang, and Ji-Rong Wen. Recbole : Towards a unified, comprehensive and efficient framework for recommendation algorithms. *CoRR*, abs/2011.01731, 2020a. URL https://arxiv.org/abs/2011.01731.

Jesus Bobadilla, Santiago Alonso, and Antonio Hernando. Deep learning architecture for collaborative filtering recommender systems. *Applied Sciences*, 10(7), 2020. ISSN 2076-3417. doi : 10.3390/app10072441. URL https://www.mdpi.com/2076-3417/10/7/2441.

Yung-Cheol Byun Zeinab Shahbazi. Product recommendation based on content-based filtering using xgboost classifier. *International Journal of Advanced Science and Technology*, 29(04) :6979 6988, Jun. 2020. URL http://sersc.org/journals/index.php/IJAST/article/view/28099.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. *CoRR*, abs/1708.05031, 2017. URL http://arxiv.org/abs/1708.05031.

Shuai Zhang, Lina Yao, and Aixin Sun. Deep learning based recommender system : A survey and new perspectives. *CoRR*, abs/1707.07435, 2017. URL http://arxiv.org/abs/1707.07435.

Feng Xue, Xiangnan He, Xiang Wang, Jiandong Xu, Kai Liu, and Richang Hong. Deep item-based collaborative filtering for top-n recommendation. *CoRR*, abs/1811.04392, 2018. URL http://arxiv.org/abs/1811.04392.

Steffen Rendle, Walid Krichene, Li Zhang, and John R. Anderson. Neural collaborative filtering vs. matrix factorization revisited. *Fourteenth ACM Conference on Recommender Systems*, 2020. URL http://proceedings.mlr.press/v139/xu21d/xu21d.pdf.

Vito Walter Anelli, Alejandro Bellogín, Tommaso Di Noia, and Claudio Pomo. Reenvisioning collaborative filtering vs matrix factorization. *CoRR*, abs/2107.13472, 2021. URL https://arxiv.org/pdf/2107.13472.pdf.

Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. Rethinking neural vs. matrix-factorization collaborative filtering : the theoretical perspectives. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11514–11524. PMLR, 18–24 Jul 2021. URL http://proceedings.mlr.press/v139/xu21d/xu21d.pdf.

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. Wide & deep learning for recommender systems. *CoRR*, abs/1606.07792, 2016. URL https://arxiv.org/pdf/1606.07792.pdf.

Rishabh Jain and Pranava Madhyastha. Model explanations under calibration. *CoRR*, abs/1906.07622, 2019. URL http://arxiv.org/abs/1906.07622.

Charu C. Aggarwal. *Recommender Systems : The Textbook*. Springer, 1st edition edition, 2016.

Cheng Guo and Felix Berkhahn. Entity embeddings of categorical variables. *CoRR*, abs/1604.06737, 2016. URL http://arxiv.org/abs/1604.06737.

Patricio Cerda, Gaël Varoquaux, and Balázs Kégl. Similarity encoding for learning with dirty categorical variables. *CoRR*, abs/1806.00979, 2018. URL http://arxiv.org/abs/1806.00979.

Diana Ferreira, Sofia Silva, António Abelha, and José Machado. Recommendation system using autoencoders. *Applied Sciences*, 10(16), 2020. ISSN 2076-3417. doi : 10.3390/app10165510. URL https://www.mdpi.com/2076-3417/10/16/5510.

Pegah Sagheb Haghighi, Olurotimi Seton, and Olfa Nasraoui. An explainable autoencoder for collaborative filtering recommendation. *CoRR*, abs/2001.04344, 2020. URL https://arxiv.org/abs/2001.04344.

Ziwei Zhu, Jianling Wang, and James Caverlee. Improving top-k recommendation via joint collaborative autoencoders. *The World Wide Web Conference*, 2019. URL https://people.engr.tamu.edu/caverlee/pubs/zhu19www.pdf.

Diederik P. Kingma and Jimmy Ba. Adam : A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

François Chollet. *Deep Learning with Python*. Manning, November 2017. ISBN 9781617294433.

Wenchuan Shi, Liejun Wang, and Jiwei Qin. User embedding for rating prediction in svd++-based collaborative filtering. *Symmetry*, 12(121), 2019. URL https://www.researchgate.net/publication/338475810_User_Embedding_for_Rating_Prediction_in_SVD-Based_Collaborative_Filtering/fulltext/5e16a52492851c8364bd486b/User-Embedding-for-Rating-Prediction-in-SVD-Based-Collaborative-Filtering.pdf.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm : A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems. Proceedings of NIPS 2017*, 30 :3149–3157, 2017. URL https://papers.nips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

Sandra Rizkallah, Amir F. Atiya, and Samir Shaheen. New vector-space embeddings for recommender systems. *Applied Sciences*, 11(14), 2021. ISSN 2076-3417. doi : 10.3390/app11146477. URL https://www.mdpi.com/2076-3417/11/14/6477/pdf.

Chuanchuan Zhao, Jinguo You, Xinxian Wen, and Xiaowu Li. Deep bi-lstm networks for sequential recommendation. *Entropy*, 22 :870, 08 2020b. doi : 10.3390/e22080870. URL https://pdfs.semanticscholar.org/c30e/300a62000d7995a9aab35671dec3d39b04f9.pdf.

Hui Fang, Danning Zhang, Yiheng Shu, and Guibing Guo. Deep learning for sequential recommendation : Algorithms, influential factors, and evaluations. *ACM Trans. Inf. Syst.*, 39(1), November 2020. ISSN 1046-8188. doi : 10.1145/3426723. URL https://arxiv.org/pdf/1905.01997.pdf.

Massimo Quadrana, Paolo Cremonesi, and Dietmar Jannach. Sequence-aware recommender systems. *CoRR*, abs/1802.08452, 2018. URL https://arxiv.org/pdf/1802.08452.pdf.

Mihaela Brut and F. Sèdes. Modélisation basée sur ontologies pour développer des recommandations personnalisées dans les systèmes hypermédia adaptatives. In *INFORSID*, 2010.

Rudi Studer, V.Richard Benjamins, and Dieter Fensel. Knowledge engineering : Principles and methods. *Data & Knowledge Engineering*, 25(1) :161–197, 1998. ISSN 0169-023X. doi : https://doi.org/10.1016/S0169-023X(97)00056-6. URL https://www.sciencedirect.com/science/article/pii/S0169023X97000566.

Thomas Effland and Michael Collins. Partially supervised named entity recognition via the expected entity ratio loss. *ArXiv*, abs/2108.07216, 2021. URL https://arxiv.org/pdf/2108.07216.pdf.

Tru Hoang Cao, Vuong M. Ngo, Dung T. Hong, and Tho T. Quan. Semantic document clustering on named entity features. *CoRR*, abs/1807.07777, 2018. URL http://arxiv.org/abs/1807.07777.

Omid Jafari, Parth Nagarkar, Bhagwan Thatte, and Carl Ingram. Satellitener : An effective named entity recognition model for the satellite domain. pages 100–107, 01 2020. doi : 10.5220/0010147401000107. URL https://www.scitepress.org/Papers/2020/101474/101474.pdf.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT : pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Nils Reimers and Iryna Gurevych. Sentence-bert : Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084, 2019. URL http://arxiv.org/abs/1908.10084.

Thomas Wolf et al. Transformers : State-of-the-art natural language processing. In Association for Computational Linguistics, editor, *Actes de la conférence 2020 sur les méthodes empiriques dans le traitement du langage naturel : démonstrations de systèmes*, pages 38–45, Online, October 2020. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning - A comprehensive evaluation of the good, the bad and the ugly. *CoRR*, abs/1707.00600, 2017. URL http://arxiv.org/abs/1707.00600.

Hui Liu, Danqing Zhang, Bing Yin, and Xiaodan Zhu. Improving pretrained models for zero-shot multi-label text classification through reinforced label hierarchy reasoning. *ArXiv*, abs/2104.01666, 2021. URL https://arxiv.org/pdf/2104.01666.pdf.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/N18-1101.

Pablo del Moral, Sawomir Nowaczyk, A. Sant'Anna, and Sepideh Pashami. Pitfalls of assessing extracted hierarchies for multi-class classification. *ArXiv*, abs/2101.11095, 2021. URL https://arxiv.org/pdf/2101.11095.pdf.

Xuepeng Wang, Li Zhao, Bing Liu, Tao Chen, Feng Zhang, and Di Wang. Concept-based label embedding via dynamic routing for hierarchical text classification. In *ACL/IJCNLP*, 2021. URL https://www.researchgate.net/publication/319968747_HDLTex_Hierarchical_Deep_Learning_for_Text_Classification.

Kamran Kowsari, Donald Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew Gerber, and Laura Barnes. Hdltex : Hierarchical deep learning for text classification. 12 2017. doi : 10.1109/ICMLA.2017.0-134.

Annexe A

Investor targeting

1 Outline

This appendix describes the overall process of the ultimate Python script that generate a list of potential investors. The data is scrapped from pitchbook.com. The model is based on a gradient boosting tree algorithm. We use the LightGBM library. It is a version of the gradient boosting method that uses leaf subsampling allowing then better computer performances. After presenting the code structure and basic command to run the code, we analyze each part of the code (initialization, data pre-processing, model training and investor prediction).

1.1 Code structure

- **main.py :** main script to run a job either in 'train' or 'predict' mode.
- **py/job.py :** The Job class handles all the meta parameters for a given model training or deal prediction.
- **py/pitchbook.py :** The PitchbookLoader class loads the pitchbook json data files, timestamps the deals, creates deals-id, corrects the investor names, checks the missing items, cleans up the duplicates and ouputs a dataframe.
- **py/model.py :** calls data.py/preprocess
- **py/data.py :** preprocesses deals and investors (new feature generation, extracts city, state, fills missing values, formats series, . . .)
- **py/transformers.py :** contains **different classes that are called in the model pipeline.** They have different goals such as :
  - column selection (**ColumnsSelector**)
  - creation of **new features** (**Featurer**),
  - **accounts for deal data temporality** by counting the occurrence of different features over the last year (**DealCounter**), or
  - counting the number of days since last deal on a specific feature (**DaysSinceLastDealCounter**), and
  - categorical data numerical **encoding** (**SelectiveNumericalEncoder**) from aikit.transformers.NumericalEncoder. The option 'num' is used and corresponds to a simple numerical encoding where each modality is transformed into a number.

- **py/utils.py :** basic function such as **expand_grid** that merge two databases, **chuncks** that truncate size for a specific dataframe column (it is not actually used), **load_json**, **load_pkl**, **save_pkl**, **get_state_mapping**, **get_features_mapping**, **get_last_model_name**, **cross_join**, **to_native_type**, **to_native_list**

- **py/metrics.py : map_eval** returns lgbm average precision score (it is not actually used), **metric_by_deal** computes metric by deal_id, **mean_recall** computes the average of all individual **deal recall** computed from the prediction results as the **ratio of predicted_investors versus actual_investors**.

1.2 Run a job

1.2.1 Environment variables

MODEL_PATH = %saved .pkl - model output directory%
PITCHBOOK_PATH = %pitchbook json data directory%

1.2.2 Training a model

```
> python main.py --mode train
```

1.2.3 Use-case prediction

```
> python main.py --mode predict --config my_deal_config.json
```

---

2 Training

The first part of the training is dedicated to load and preprocess the pitchbook data (deals.json + investors.json).

```
data = PitchbookLoader(job.data_path)
data.load()
deals       = preprocess_deals(data.deals)
investors   = preprocess_investors(data.investors)
```

2.1 Loading the Pitchbook data

The 'train' mode from 'main.py' allows to load directly the .json pitchbook data without installing the pitchbook scrapping script. The dependency from the scrapping pitchbook module is then removed by the use of the class **PitchbookLoader** from *pitchbook.py*. NB : set the environment variable PITCHBOOK_PATH to the path with the .json pitchbook data.

The loading class **PitchbookLoader** from *pitchbook.py* completes the following tasks :

- loading all json files. NB : Only investors.json and deals.json are actually used.
- creating *deal['deal_id']* feature by associating the company id to each timestamped deal.
- correcting investor names (for each *deal_id*, we replace the investor names in *deals* by the ones in *investors*).
- removing duplicated deals
- running all the functions that clean up the database by removing the duplicates and checking the missing items (cf **PitchbookLoader().sanity_check**).

2.2 Deal preprocessing

From *data.py*, the function **preprocess_deals** completes the following tasks :

- select completed deals only
- choose deal features (columns selection). Currently selecting *deal_id*, *deal_date*, *deal_type*, *series*, *vc_round*, *deal_size*, *percent_acquired*, *company_id*, *company_name*, *primary_industry_sector*, *primary_industry_group*, *primary_industry_code*, *hq_location*, *verticals*
- format deal *deal_date* and sort deals by *deal_date*
- format deal *series* by applying the function **process_series**
  - missing values are set NaN
  - only keeps the first letter describing the serie. Correct final values are either NaN or ['A', 'B', 'C', 'D', 'E', 'F']
- format deal *vc_round* by applying the function **process_vc_round**
  - fill with NaN null or empty string values, and numeric vc round input
- format deal *verticals* by applying the function **process_verticals**
  - fill with NaN null or empty string values
  - split vertical input string and only take the first vertical word
- rename *hq_location* column by *company_hq_location*
- for each categorical feature, fill missing values by *'Unknown'* by applying the function **fill_missing_object_values**

2.3 Investor preprocessing

From *data.py*, the function **preprocess_investors** complete the following tasks :

- choose investors features (columns selection). Currently selecting *investor_id*, *investor_name*, *primary_investor_type*, *aum*, *year_founded*, *hq_location*
- rename *hq_location* column by *investor_hq_location*
- create the investor feature *investor_hq_city* from investor *investor_hq_location* by applying the function **extract_city**
- create the investor feature *investor_hq_state* from investor *investor_hq_location* by applying the function **extract_state**
  NB : both functions (extract_city and extract_state) require the location to be formatted as 2 separate words such as *'city, state'*.
- create the investor feature *investor_hq_country* from investor *investor_hq_state* by applying the function **get_state_mapping('country')** from *utils.py*
- create the investor feature *investor_hq_region* from investor *investor_hq_state* by applying the function **get_state_mapping('region')** from *utils.py*
  The geographical mapping function uses the *'parameters.json'* file that delivers :
  - the correspondance between the countries or states and the (country, region)
  - the feature name and its description
    (used by the function **utils.get_features_mapping**)
- for each categorical feature, fill missing values by *'Unknown'* by applying the function **fill_missing_object_values**

Next, we analyze the second part of the training :

```
model = InvestorTargeter()
model.fit(deals, investors, data.deals_investors)
save_pkl(model, os.path.join(job.models_path, model.get_name()))
```

The second part of the training is then dedicated to :

- compute the model pipeline
- save the trained model

2.4 InvestorTargeter

The model class **InvestorTargeter** from *model.py* is loaded. It completes the following main tasks :

1. initialization
2. fit
3. predict

2.4.1 InvestorTargeter inititalization

- **InvestorTargeter().init —- parameter definition and initialization :**
  - *min_number_of_investments :* deals with active_investors less than this value are dropped
  - *downsample_ratio :* if set at 0.02, 2% of the dataset is randomly selected for training the model (cf **InvestorTargeter()._preprocess**). NB : While the training data *X_train* is downsampled, the validation data *X_valid* is not.
  - *confidence_adjustment_ratio :* Adjustment ratio to manually rescale probabilities between 0 and 1 (cf **InvestorTargeter()._upsample_prediction**)
  - *validation_ratio :* if set at 0.01, 99% of the deals are used for training (cf **InvestorTargeter()._extract_training_and_validation_deals**)
  - *model_id :* Model unique identifier
  - *last_deal_date :* Max deal date (cf **fit._extract_inputs_information**)
  - *training_deals :* train dataset, wil be called *X_train* after preprocessing (cf **InvestorTargeter()._preprocess**)
  - *validation_deals :* validation dataset, will be called *X_valid* after preprocessing (cf **InvestorTargeter()._preprocess**)
  - *deals :* Dataframe of deals
  - *investors :* Dataframe of investors
  - *investments :* Dataframe on selected deals and known investors. The key ['deal_id', 'investor_id'] is used to initialize the matching investments and label investments['invest'] = 1 (cf **InvestorTargeter()._extract_data_on_active_investors** and **InvestorTargeter()._preprocess**)
  - *pipeline :* model fit pipeline including new feature generation and column selection (cf **InvestorTargeter()._preprocess**)
  - *model :* LightGBM model
  - *scores :* Model scores
  - *explainer :* Shap library (*i.e.* feature explainer)
  - *categories :* kept categorical columns to be numerically encoded (cf aikit.transformers **NumericalEncoder** in **SelectiveNumericalEncoder** from *transformers.py*)
    - **deal description :** 'deal_type', 'series', 'vc_round',
    - **company industry sector :** 'primary_industry_sector', 'primary_industry_group', 'primary_industry_code', 'verticals',
    - **company geographical sector :** 'company_hq_city', 'company_hq_state', 'company_hq_country', 'company_hq_region',

- **investor description :** 'primary_investor_type',
- **investor geographical sector :** 'investor_hq_city', 'investor_hq_state', 'investor_hq_country', 'investor_hq_region'

- *params :* LightGBM parameters (NB : No hyperparameter tuning implemented so far !)
- *inputs :* Selection of input features to make prediction
  - **deal description :** 'deal_type', 'series', 'vc_round', 'deal_size', 'percent_acquired',
  - **company description (industry and geographical sector) :** 'company_id', 'primary_industry_sector', 'primary_industry_group', 'primary_industry_code', 'verticals', 'company_hq_location'

2.4.2 InvestorTargeter fit

The **fit** function in the **InvestorTargeter** class is the hardcore of the program with the model pipeline.

```
def fit(self, deals, investors, investments):
   self._init()
   self._extract_inputs_information(deals)
   self._extract_data_on_active_investors(deals, investors, investments)
   self._extract_training_and_validation_deals()

   X_train = self._preprocess(self.training_deals, sample=True)
   X_valid = self._preprocess(self.validation_deals, sample=False)
```

- **__init :** Timestamp the model id (cf **get__name**) so that we have a unique model identifier, initialize the scores
- **__extract__inputs__information :** Set max/last deal date. Extract list of values for each input. Add specific value for unknown company.
- **__extract__data__on__active__investors :** Compute *active_investors* as the number of investments for each investor_id
- **__extract__training__and__validation__deals :** Split deals data using validation_ratio. Dump variables fit.training_deals and fit.validation_deals
- **__preprocess :** Preprocess *X_train* and *X_valid*. Prepare the data for model application (eventual subsampling, label initialization, monitor positive ratio). NB : While the training data *X_train* is downsampled, the validation data *X_valid* is not.
- X dataframe creation : here we concatenate and merge train and validation data and sort values by ['investor_id', 'deal_date']
- computes new feature X['investor_age'] to filter out not yet founded investors
- define the model pipeline :

```
self.pipeline = make_pipeline(
    Featurer(),
    DealCounter(),
    DealCounter(['deal_type']),
    DealCounter(['series']),
    DealCounter(['vc_round']),
    DealCounter(['company_id']),
    DealCounter(['primary_industry_sector']),
    DealCounter(['primary_industry_group']),
    DealCounter(['primary_industry_code']),
    DealCounter(['company_hq_country']),
```

```
            DealCounter(['company_hq_region']),
            DealCounter(['verticals']),
            DaysSinceLastDealCounter(),
            DaysSinceLastDealCounter(['deal_type']),
            DaysSinceLastDealCounter(['primary_industry_sector']),
            DaysSinceLastDealCounter(['primary_industry_group']),
            DaysSinceLastDealCounter(['primary_industry_code']),
            DaysSinceLastDealCounter(['company_hq_region']),
            SelectiveNumericalEncoder(self.categories),
            ColumnsSelector(columns_to_drop=[
                'deal_id', 'deal_date', 'invest',
                'company_id', 'company_name',
                'investor_id', 'investor_name', 'investor_hq_location'
            ])
        )
```

NB : In perspective, the data preprocessing contained in this function should be moved out into functions as the **InvestorTargeter()._preprocessing** is. For example, all the data from the first year is filtered out, as features are not properly computed yet (1 year of data required for the 'rolling' features).

- **__train__model :** The LightGBM model is computed by using lgb.train on the train_set and valid_set

```
X_train, y_train = X[training], y[training]
X_valid, y_valid = X[validation], y[validation]

train_set = lgb.Dataset(X_train, y_train)
valid_set = lgb.Dataset(X_valid, y_valid)

self.model = lgb.train(
    self.params,
    train_set,
    valid_sets=[valid_set],
    num_boost_round=10000,
    early_stopping_rounds=50,
    verbose_eval=50
)
```

- **__compute__metrics :** roc_auc_score, average_precision_score. These metrics are also computed by deal (metric_by_deal on roc_auc_score,
metric_by_deal on average_precision_score). mean_recall

2.4.3 InvestorTargeter predict

- **predict**
  - convert json deal and company description inputs to dataframe
  - **__validate__inputs :** input format validation
  - merge input prediction config dataframe with investor dataframe (cf expand_grid from utils.py)
  - initialize investment score : X['invest'] = 0
  - apply **pipeline.transform** to dataframe and fill prediction investment score results : X['invest'] = self.model.predict(self.pipeline.transform(X))

- probability distribution resampling (cf **InvestorTargeter()._upsample_prediction**)
- investment score sorting and selection of a list of the first *number_of_investors* investors
- return result dataframe on selected columns : 'investor_id', 'investor_name', 'primary_investor_type', 'aum', 'investor_hq_location', 'investor_hq_country', 'investor_hq_region', 'probability'
- calls ExplanationFactory() from model.py
  (cf **InvestorTargeter()._generate_results**). Process shap library feature explanation. SHAP (SHapley Additive exPlanations) quantifies the contribution that each feature brings to the prediction made by the model. To do so, SHAP computes what is called the "marginal contribution" brought by each feature to the model.

---

3 Prediction

3.1 Loading the parameters

The prediction of potential investors for a specific deal is described here. First, the deal configuration is loaded : the .json config parameter file is read by using the class **Job** defined in *job.py*.

```
deal = job.config["deal"]
```

Then, the .pkl model file is loaded. Finally, the predict function of the model is applied to the deal input parameters to output the potential investor list.

```
model: InvestorTargeter = load_pkl(os.path.join(job.models_path, \
job.config['model_filename']))
investors = model.predict(deal)
```

4 Sample of log file related to a transaction for Ynsect

```
possible_series:  ['C']
possible_deal_type:  ['Later Stage VC']
possible_primary_industry_sector:  ['Business Products and Services (B2B)',
'Materials and Resources']
possible_primary_industry_group:  ['Agriculture', 'Commercial Products']
possible_verticals:  ['AgTech', 'CleanTech', 'LOHAS & Wellness']
possible_primary_industry_code: ['Other Agriculture']
possible_company_hq_location: ['Paris, France', 'Unknown']
{
"auc": 0.9086732663936897,
"ap": 0.03777713041899044,
"mean_auc": 0.8968878017627465,
"mean_ap": 0.08711945391078678,
"mean_recall": 0.35880756393090363
}
======================================================================
Later Stage VC C Agriculture Business Products and Services (B2B)
Other Agriculture AgTech Paris, France
====================
{
"deal_type": "Later Stage VC",
"series": "C",
```

```
"vc_round": "5th",
"deal_size": 275,
"percent_acquired": 0.5,
"company_id": "Unknown",
"primary_industry_sector": "Business Products and Services (B2B)",
"primary_industry_group": "Agriculture",
"primary_industry_code": "Other Agriculture",
"company_hq_location": "Paris, France",
"verticals": "AgTech"
}
features:  Index(['deal_type', 'series', 'vc_round', 'deal_size', 'percent_acquired',
'company_id', 'primary_industry_sector', 'primary_industry_group',
'primary_industry_code', 'company_hq_location', 'verticals',
'deal_date', 'deal_id', 'company_name', 'investor_id', 'investor_name',
'primary_investor_type', 'aum', 'year_founded', 'investor_hq_location',
'investor_hq_city', 'investor_hq_state', 'investor_hq_country',
'investor_hq_region', 'invest', 'company_hq_city', 'company_hq_state',
'company_hq_country', 'company_hq_region', 'probability'],
dtype='object')
[
{
"investor_id": "110379-70",
"investor_name": "Armat Group",
"primary_investor_type": "PE/Buyout",
"aum": NaN,
"investor_hq_location": "Luxembourg, Luxembourg",
"investor_hq_country": "Luxembourg",
"investor_hq_region": "Europe",
"probability": 0.13401385110009909,
"explanation": [
{
"feature": "Number of deals last year",
"value": 1.0,
"influence": 3.735725397730675
},
{
"feature": "Company country",
"value": "France",
"influence": 1.1319855203634788
},
{
"feature": "Number of deals on country last year",
"value": 1.0,
"influence": 0.9611841268033242
},
{
"feature": "Number of deals of series last year",
"value": 1.0,
"influence": -0.8822957517945887
},
{
"feature": "Days since last deal on region last year",
"value": 65.0,
"influence": 0.8729619810722138
```

```
}
]
},
{
"investor_id": "59107-42",
"investor_name": "Compagnie du Bois Sauvage (BRU: COMB)",
"primary_investor_type": "Corporation",
"aum": NaN,
"investor_hq_location": "Brussels, Belgium",
"investor_hq_country": "Belgium",
"investor_hq_region": "Europe",
"probability": 0.11896059612523226,
"explanation": [
{
"feature": "Number of deals last year",
"value": 2.0,
"influence": 3.164125135960043
},
{
"feature": "Company country",
"value": "France",
"influence": 1.0888692304279293
},
{
"feature": "Number of deals on country last year",
"value": 1.0,
"influence": 1.0732829971249818
},
{
"feature": "Days since last deal on region last year",
"value": 65.0,
"influence": 0.892725982916803
},
{
"feature": "Days since last deal on deal type last year",
"value": 65.0,
"influence": 0.839567000668373
}
]
},
{
"investor_id": "42482-26",
"investor_name": "Bouygues (PAR: EN)",
"primary_investor_type": "Corporation",
"aum": NaN,
"investor_hq_location": "Paris, France",
"investor_hq_country": "France",
"investor_hq_region": "Europe",
"probability": 0.10861329899703674,
"explanation": [
{
"feature": "Number of deals last year",
"value": 1.0,
"influence": 87.46806299041579
```

```
},
{
"feature": "Days since last deal on deal type last year",
"value": NaN,
"influence": -25.99907022346812
},
{
"feature": "Days since last deal on region last year",
"value": 365.0,
"influence": -22.30151621974209
},
{
"feature": "Number of deals of series last year",
"value": 0.0,
"influence": -20.99949618855988
},
{
"feature": "n_deals_primary_industry_code",
"value": 0.0,
"influence": -3.525280654262344
}
]
},
{
"investor_id": "14141-98",
"investor_name": "Caisse des D\u00e9p\u00f4ts Group",
"primary_investor_type": "Sovereign Wealth Fund",
"aum": 174962.67,
"investor_hq_location": "Paris, France",
"investor_hq_country": "France",
"investor_hq_region": "Europe",
"probability": 0.09759802477268531,
"explanation": [
{
"feature": "Number of deals last year",
"value": 3.0,
"influence": 2.7145371924320596
},
{
"feature": "Number of deals on country last year",
"value": 3.0,
"influence": 1.4275310677786721
},
{
"feature": "Company country",
"value": "France",
"influence": 1.4189726468181207
},
{
"feature": "Days since last deal on region last year",
"value": 65.0,
"influence": 1.1502385420034744
},
{
```

```
"feature": "Days since last deal on deal type last year",
"value": 65.0,
"influence": 1.1027023401444365
}
]
},
{
"investor_id": "265675-15",
"investor_name": "Astanor Ventures",
"primary_investor_type": "Impact Investing",
"aum": 329.92,
"investor_hq_location": "London, United Kingdom",
"investor_hq_country": "UK",
"investor_hq_region": "Europe",
"probability": 0.09714961239509787,
"explanation": [
{
"feature": "Number of deals last year",
"value": 7.0,
"influence": 1.546491616551045
},
{
"feature": "Number of deals on verticals last year",
"value": 6.0,
"influence": 1.445105454982487
},
{
"feature": "Days since last deal on region last year",
"value": 9.0,
"influence": 0.89332886310309
},
{
"feature": "Number of deals on country last year",
"value": 1.0,
"influence": 0.8701942847831555
},
{
"feature": "Number of deals of series last year",
"value": 2.0,
"influence": -0.8142840701754819
}
]
},
{
"investor_id": "10597-15",
"investor_name": "Unigrains",
"primary_investor_type": "Growth/Expansion",
"aum": 892.53,
"investor_hq_location": "Paris, France",
"investor_hq_country": "France",
"investor_hq_region": "Europe",
"probability": 0.08701321385866037,
"explanation": [
{
```

```
"feature": "Number of deals last year",
"value": 1.0,
"influence": 19.7183478460701
},
{
"feature": "Number of deals of series last year",
"value": 0.0,
"influence": -3.692686239307523
},
{
"feature": "Days since last deal on region last year",
"value": 318.0,
"influence": -3.5282749859384843
},
{
"feature": "Days since last deal on deal type last year",
"value": 318.0,
"influence": -2.993237203915027
},
{
"feature": "Investor country",
"value": "France",
"influence": 2.657338694662969
}
]
},
{
"investor_id": "166283-74",
"investor_name": "Supernova Invest",
"primary_investor_type": "Asset Manager",
"aum": 274.13,
"investor_hq_location": "Paris, France",
"investor_hq_country": "France",
"investor_hq_region": "Europe",
"probability": 0.08014306299916832,
"explanation": [
{
"feature": "Number of deals on country last year",
"value": 7.0,
"influence": 1.6545308061531296
},
{
"feature": "Number of deals last year",
"value": 7.0,
"influence": 1.4305777598064764
},
{
"feature": "Company country",
"value": "France",
"influence": 1.3823186876859015
},
{
"feature": "Days since last deal on region last year",
"value": 14.0,
```

```
"influence": 0.9632348360639873
},
{
"feature": "Number of deals of series last year",
"value": 2.0,
"influence": -0.8395667136226364
}
]
},
{
"investor_id": "61488-10",
"investor_name": "Aucfan (TKS: 3674)",
"primary_investor_type": "Corporation",
"aum": NaN,
"investor_hq_location": "Tokyo, Japan",
"investor_hq_country": "Japan",
"investor_hq_region": "APAC",
"probability": 0.06827199695467007,
"explanation": [
{
"feature": "Deal size (m$)",
"value": 275.0,
"influence": 7.287587443035464
},
{
"feature": "Number of deals of deal type last year",
"value": 1.0,
"influence": 6.690344374176241
},
{
"feature": "Days since last deal on deal type last year",
"value": 197.0,
"influence": -4.948436779050646
},
{
"feature": "Number of deals last year",
"value": 2.0,
"influence": 4.684555156892172
},
{
"feature": "Number of deals of series last year",
"value": 0.0,
"influence": -3.5363783844216026
}
]
},
{
"investor_id": "42269-77",
"investor_name": "Cr\u00e9dit Mutuel Ark\u00e9a",
"primary_investor_type": "Corporation",
"aum": 37004.18,
"investor_hq_location": "Le Relecq-Kerhuon, France",
"investor_hq_country": "France",
"investor_hq_region": "Europe",
```

```
"probability": 0.0503028748982409,
"explanation": [
{
"feature": "Number of deals last year",
"value": 2.0,
"influence": 2.8857405955349815
},
{
"feature": "Company country",
"value": "France",
"influence": 1.4434776196870343
},
{
"feature": "Days since last deal on region last year",
"value": 91.0,
"influence": 1.3360231845723953
},
{
"feature": "Number of deals on country last year",
"value": 2.0,
"influence": 1.1864072794632325
},
{
"feature": "Days since last deal on deal type last year",
"value": 91.0,
"influence": 1.0649522965922946
}
]
},
{
"investor_id": "51056-11",
"investor_name": "Creadev",
"primary_investor_type": "Family Office",
"aum": NaN,
"investor_hq_location": "Paris, France",
"investor_hq_country": "France",
"investor_hq_region": "Europe",
"probability": 0.05030056541533191,
"explanation": [
{
"feature": "Number of deals last year",
"value": 5.0,
"influence": 2.111862362115549
},
{
"feature": "Company country",
"value": "France",
"influence": 1.428287248380847
},
{
"feature": "Days since last deal on region last year",
"value": 15.0,
"influence": 1.4110624115712789
},
```

```
{
"feature": "Days since last deal on deal type last year",
"value": 23.0,
"influence": 1.2157573117703444
},
{
"feature": "Number of deals of series last year",
"value": 0.0,
"influence": -1.1174185019565124
}
]
},
{
"investor_id": "56595-07",
"investor_name": "Ingka GreenTech",
"primary_investor_type": "Corporate Venture Capital",
"aum": 58.29,
"investor_hq_location": "Malm\u00f6, Sweden",
"investor_hq_country": "Sweden",
"investor_hq_region": "Europe",
"probability": 0.04735811121048564,
"explanation": [
{
"feature": "Number of deals last year",
"value": 2.0,
"influence": 3.2065160100253687
},
{
"feature": "Company country",
"value": "France",
"influence": 1.2718742712274822
},
{
"feature": "Days since last deal on region last year",
"value": 64.0,
"influence": 1.053689716351767
},
{
"feature": "Number of deals on country last year",
"value": 1.0,
"influence": 1.0209613034680143
},
{
"feature": "Number of deals of series last year",
"value": 0.0,
"influence": -0.9898009668924034
}
]
},
{
"investor_id": "229204-18",
"investor_name": "Welcan Capital",
"primary_investor_type": "Venture Capital",
"aum": NaN,
```

```
"investor_hq_location": "Clark, NJ",
"investor_hq_country": "USA",
"investor_hq_region": "Northern America",
"probability": 0.04423512493966181,
"explanation": [
{
"feature": "Number of deals last year",
"value": 1.0,
"influence": 79.4737263045321
},
{
"feature": "Days since last deal on deal type last year",
"value": NaN,
"influence": -27.571423418406425
},
{
"feature": "Days since last deal on region last year",
"value": NaN,
"influence": -23.20795150885124
},
{
"feature": "Number of deals of series last year",
"value": 0.0,
"influence": -21.796180752552107
},
{
"feature": "Number of deals of deal type last year",
"value": 0.0,
"influence": 6.066633522408527
}
]
},
{
"investor_id": "232649-83",
"investor_name": "AFI Capital Partners",
"primary_investor_type": "Venture Capital",
"aum": NaN,
"investor_hq_location": "Seattle, WA",
"investor_hq_country": "USA",
"investor_hq_region": "Northern America",
"probability": 0.04423512493966181,
"explanation": [
{
"feature": "Number of deals last year",
"value": 1.0,
"influence": 79.4737263045321
},
{
"feature": "Days since last deal on deal type last year",
"value": NaN,
"influence": -27.571423418406425
},
{
"feature": "Days since last deal on region last year",
```

```
"value": NaN,
"influence": -23.20795150885124
},
{
"feature": "Number of deals of series last year",
"value": 0.0,
"influence": -21.796180752552107
},
{
"feature": "Number of deals of deal type last year",
"value": 0.0,
"influence": 6.066633522408527
}
]
},
{
"investor_id": "100564-30",
"investor_name": "New Protein Capital",
"primary_investor_type": "Venture Capital",
"aum": 40.0,
"investor_hq_location": "Singapore, Singapore",
"investor_hq_country": "Singapore",
"investor_hq_region": "APAC",
"probability": 0.042917040190008166,
"explanation": [
{
"feature": "Number of deals last year",
"value": 2.0,
"influence": 2.9800418929019377
},
{
"feature": "Company country",
"value": "France",
"influence": 1.0768417621649915
},
{
"feature": "Number of deals on country last year",
"value": 1.0,
"influence": 1.0315223639422866
},
{
"feature": "Days since last deal on region last year",
"value": 65.0,
"influence": 0.9373736655050602
},
{
"feature": "Number of deals of series last year",
"value": 1.0,
"influence": -0.8980221969568836
}
]
},
{
"investor_id": "225660-25",
```

```
"investor_name": "RATP Capital Innovation",
"primary_investor_type": "Corporation",
"aum": NaN,
"investor_hq_location": "Paris, France",
"investor_hq_country": "France",
"investor_hq_region": "Europe",
"probability": 0.04195173895124836,
"explanation": [
{
"feature": "Number of deals last year",
"value": 1.0,
"influence": 15.270480188995467
},
{
"feature": "Days since last deal on region last year",
"value": 350.0,
"influence": -6.399128228562882
},
{
"feature": "Number of deals of series last year",
"value": 0.0,
"influence": -3.721460900513331
},
{
"feature": "Investor country",
"value": "France",
"influence": 2.8238098706236117
},
{
"feature": "Number of deals of VC round last year",
"value": 0.0,
"influence": -2.19691829341971
}
]
},
{
"investor_id": "53596-36",
"investor_name": "Valeo (PAR: FR)",
"primary_investor_type": "Corporation",
"aum": 528.75,
"investor_hq_location": "Paris, France",
"investor_hq_country": "France",
"investor_hq_region": "Europe",
"probability": 0.041486573796228055,
"explanation": [
{
"feature": "Number of deals last year",
"value": 1.0,
"influence": 3.4026530147479606
},
{
"feature": "Company country",
"value": "France",
"influence": 1.5026817009575493
```

```
},
{
"feature": "Days since last deal on region last year",
"value": 64.0,
"influence": 1.076001937064779
},
{
"feature": "Number of deals on country last year",
"value": 1.0,
"influence": 0.9536265524550869
},
{
"feature": "Number of deals of series last year",
"value": 0.0,
"influence": -0.9185258517396653
}
]
},
{
"investor_id": "97655-68",
"investor_name": "Capagro",
"primary_investor_type": "Venture Capital",
"aum": 147.79,
"investor_hq_location": "Paris, France",
"investor_hq_country": "France",
"investor_hq_region": "Europe",
"probability": 0.04031441562387311,
"explanation": [
{
"feature": "Number of deals last year",
"value": 4.0,
"influence": 2.145143852148277
},
{
"feature": "Number of deals on country last year",
"value": 4.0,
"influence": 1.6001967432650106
},
{
"feature": "Company country",
"value": "France",
"influence": 1.457458928297231
},
{
"feature": "Number of deals of series last year",
"value": 0.0,
"influence": -1.0551508906573628
},
{
"feature": "Days since last deal on region last year",
"value": 86.0,
"influence": 0.9819450102696786
}
]
```

```
},
{
"investor_id": "227190-25",
"investor_name": "Luxury Tech Fund",
"primary_investor_type": "Venture Capital",
"aum": NaN,
"investor_hq_location": "Paris, France",
"investor_hq_country": "France",
"investor_hq_region": "Europe",
"probability": 0.03719828546818616,
"explanation": [
{
"feature": "Number of deals last year",
"value": 1.0,
"influence": 4.181098257732018
},
{
"feature": "Investor country",
"value": "France",
"influence": 2.7589082413319046
},
{
"feature": "Days since last deal on deal type last year",
"value": 233.0,
"influence": -2.5278913351405974
},
{
"feature": "Deal size (m$)",
"value": 275.0,
"influence": 2.3928398545908194
},
{
"feature": "Number of deals of deal type last year",
"value": 1.0,
"influence": 2.1939508544048842
}
]
},
{
"investor_id": "52774-75",
"investor_name": "Serena Capital",
"primary_investor_type": "Venture Capital",
"aum": 365.9,
"investor_hq_location": "Paris, France",
"investor_hq_country": "France",
"investor_hq_region": "Europe",
"probability": 0.03680850861654666,
"explanation": [
{
"feature": "Company country",
"value": "France",
"influence": 1.6364208978131636
},
{
```

```
"feature": "Number of deals on country last year",
"value": 5.0,
"influence": 1.5760094452361864
},
{
"feature": "Number of deals last year",
"value": 7.0,
"influence": 1.4425438401982247
},
{
"feature": "Days since last deal on region last year",
"value": 73.0,
"influence": 0.9486307859800095
},
{
"feature": "Number of deals of series last year",
"value": 2.0,
"influence": -0.7527712574187433
}
]
},
{
"investor_id": "10028-71",
"investor_name": "BNP Paribas (PAR: BNP)",
"primary_investor_type": "Investment Bank",
"aum": 1149290.42,
"investor_hq_location": "Paris, France",
"investor_hq_country": "France",
"investor_hq_region": "Europe",
"probability": 0.036634171154428526,
"explanation": [
{
"feature": "Number of deals on region last year",
"value": 5.0,
"influence": 1.4501711561697126
},
{
"feature": "Deal size (m$)",
"value": 275.0,
"influence": 1.3860922867441996
},
{
"feature": "Investor AUM",
"value": 1149290.42,
"influence": -1.3009396047960342
},
{
"feature": "Company country",
"value": "France",
"influence": 1.1290549414233337
},
{
"feature": "Days since last deal on deal type last year",
"value": 4.0,
```

```
"influence": 1.1245834439553195
}
]
},
{
"investor_id": "81906-40",
"investor_name": "Air Liquide Venture Capital",
"primary_investor_type": "Corporate Venture Capital",
"aum": NaN,
"investor_hq_location": "Paris, France",
"investor_hq_country": "France",
"investor_hq_region": "Europe",
"probability": 0.036444642489618,
"explanation": [
{
"feature": "Number of deals last year",
"value": 2.0,
"influence": 22.006028963491147
},
{
"feature": "Days since last deal on region last year",
"value": 365.0,
"influence": -6.077884417651072
},
{
"feature": "Number of deals of series last year",
"value": 1.0,
"influence": -4.0094878350437675
},
{
"feature": "Investor country",
"value": "France",
"influence": 2.792981912483007
},
{
"feature": "Days since last deal on industry sector last year",
"value": 365.0,
"influence": -2.5340400251375086
}
]
},

.../...
```

Annexe B

Feedback classification

1 Catégories

Les catégories à plusieurs niveaux peuvent être les suivantes :

- Environment
    - Legal
        - Regulation
        - Litigation
    - Social
        - Unions
        - Regulations
    - Macro economic
        - Currency
        - Interest rates
        - Central banks
    - Political
    - Fiscal
    - Geopolitics
        - Tariffs
        - Terrorist
        - War
        - Riots
    - Crisis
        - Climate Change
        - Health crisis
        - Weather related
        - Seismic

- Industry specifics
    - Industry dynamic
    - Industry capacity
    - Disruption

- Market
    - Market share
    - Market leadership
    - Market size
    - Market exposure
    - Market growth

    * Market trend
    * Market inefficiency

- Business
    * Business model
        - Restructuring
        - Structural issue
        - Cyclicity (anything related to cycles)
        - Resilience
        - Forecast / visibility
    * Geography
        - Geographic diversification
        - Exposure
    * Customer
        - Customer acquisition
        - Customer Retention
        - Customer Churn rate
        - Online Penetration
        - Client Diversification
    * Distribution
    * Product
        - Product diversification
        - Product pricing
        - Product demand
    * Company
        - Size / scale
        - Employees / Staff / HR
    * Competition
        - Barriers to entry
        - Competitors
        - Pricing positioning

- Assets
    * Brand / awareness
    * Means of operation
        - Factory
        - Distribution network
        - supply chain
        - Logistics
    * Capital intensity
    * Intellectual property
    * Depreciation
    * Innovation
        - R&D
        - Patents
        - Technologies

- Financials
    * Liquidity (generation de cash flow)
    * Income statement
        - Profitability
        - Cost inflation
        - Margin
        - Revenues

☐ Revenue visibility : Backlog
☐ Revenue split
☐ Organic growth
☐ External growth
☐ Like for like

Tax rates

* Cash flow statement
  - Cash consumption / Burning rate
  - Capex
  - Working Capital
* Capital Structure
  - Debt
  - Leverage
  - Treasury
  - Liquid asset

- Valuation
  * Comparables
  * Return to shareholder
    - Buyback
    - Dividend policy / distribution policy
  * Return on capital
  * Return on equity
  * Equity value / Market Capitalization
  * Enterprise value
  * Share pricing / Market Value
  * Share liquidity
  * Valuation Methodology
    - Ratios : EV/EBITDA, DCF,
  * Accretion / Dilution + EPS
  * Cost of capital / Cost of equity
  * Risk premium
  * Commodities pricing
  * Interest rates

- Management
  * Leadership
  * Board
  * Execution
  * Track record

- Strategy
  * Strategy definition
  * Restructuring
  * M&A
    - Bolt on
    - Pacman
    - Integration
    - Business Expansion (vertical / horizontal)

- ESG criteria
  * 17 sustainability criteria (Sustainable Development Goals or SDGs) defined by the United Nations
  * Rating ESG

- Deal
    - ∗ Deal structure (primary / secondary)
    - ∗ Deal appetite / willingness
    - ∗ Liquidity (taille du deal)
    - ∗ Deal timing
    - ∗ Nature of seller
    - ∗ Shareholder structure
    - ∗ Shareholder dilution
    - ∗ Free float
    - ∗ Roadshow
    - ∗ Overhang
    - ∗ Capital market day

---

2 Résultats de classification de commentaires d'investisseurs

sentence	max_score	best_class
it was not that much in elis, so it is quite resilient.	0.997927	resiliency
even there, they have variable costs, they have also fixed fees coming from the hotels, so even under hospitality i see quite a good resilience.	0.997919	resiliency
the second thing is probably the resilience, except probably the hospitality part of the business.	0.997213	resiliency
all the guys running a private equity fund in paris, they know the company very well, for the last 20 years, and they know how resilient the business is.	0.997029	resiliency
they are quite resilient, so if the macro economy is a little bit harder this year or next year, we think that it is quite a defensive play, so it is a good thing.	0.996770	resiliency
the experienced management team and their established track record of executing on the business model and delivering the consolidation strategy is also referenced as a strength.	0.985960	Track record
this is what happened in 2009, this is what is happening now.	0.985224	Track record
they have proven this last quarter that they could improve the situation in the uk & ireland, so, yes, i expect they could restore growth there in 2019.	0.983766	Track record
my personal view is it probably can and it has proven historically that it can.	0.982192	Track record
he started in 1996 or something.	0.972157	Track record
bunzl is slightly better, in my opinion.	0.940191	Technology
seven, for their technical capacity.	0.907880	Technology
when they ipod, they were targeting, in terms of operating margin, 16.5% for 2017.	0.901715	Technology
they are doing bolt-ons, but they do not give out any numbers.	0.870583	Technology
they have been executing quite well in terms of integration.	0.844777	Technology
on a scale of 1 to 10 how do you rate the execution of elis strategy?	0.993997	Strategy execution
the spider graph below plots plots elis' execution of strategy rating against other companies (a to bj) that comprise the ip index (red line).	0.990437	Strategy execution
they deliver on their strategy.	0.988476	Strategy execution
they do not really need it, except in germany, but in europe now, they are leaders in all the countries - midcap investors were asked to rate elis' execution of strategy from 1 (low) to 10 (high).	0.987076	Strategy execution
execution of elis roll up, or compounding, strategy to date is widely supported and applauded.	0.982323	Strategy execution
i am talking about the social background, which will impact the economy.	0.995324	Social
maybe there could be threats in france in the very quick future because we have some social unrest in the country.	0.988603	Social
it is pretty supportive.	0.988129	Social

it is like a lost society.	0.950625	Social
it is a ten.	0.977046	Size ou Scale
one, two - hsbc four.	0.969912	Size ou Scale
four or five.	0.962607	Size ou Scale
he is really a ten.	0.961530	Size ou Scale
it is a very big acquisition.	0.959031	Size ou Scale
in this area, we are shareholders, as you may know.	0.998383	Shareholders
we are a shareholder again now.	0.996811	Shareholders
by the way, we are long-term shareholders in elis.	0.996255	Shareholders
the biggest problem at this point for shareholders is the strategy of doing other big acquisitions, or not so big, but material acquisitions with the high leverage balance sheet they have.	0.995991	Shareholders
all i can say is, we have been shareholders for a number of years and i guess that should tell you something.	0.994977	Shareholders
that is the struggle the shares have at the moment.	0.990175	Share liquidity
he bought 40,000 of shares.	0.978820	Share liquidity
we are generalists and we own a lot of shares.	0.959953	Share liquidity
i think that the berendsen deal is an opportunistic deal that, because the share price collapsed, they could take an opportunistic approach.	0.954608	Share liquidity
once the berendsen deal is past, meaning that the capex load for the next two years is done, this company is going to generate huge amounts of cash.	0.953128	Share liquidity
for the stock to do well, they probably need to do at least 3% organic revenue growth with margin expansion and people need to be comfortable with that.	0.982685	Revenues growth
not so good at organic revenue growth - dk partners i like them very much.	0.974005	Revenues growth
one would be that the receivables have gone up.	0.958488	Revenues growth
the benefit of a strong competitive position and network effect, together with synergies as bolt-on acquisitions are integrated, can drive top line growth, improve margins and also demonstrate a good roce.	0.955709	Revenues growth
not so good at organic revenue growth.	0.951408	Revenues growth
there is only a revenue number, so i do not really have much to say about it.	0.870179	Revenues KPI
it is like margin and revenue.	0.852287	Revenues KPI
people think that they are going to invest 22% of sales forever.	0.542504	Revenues KPI
as i said, the shares are all the way back to where they started, more or less.	0.993321	Return to shareholders

it is pretty much back to where i bought my shares in the first place.	0.992643	Return to shareholders
if you are looking at cash return to shareholders, there really has not been much - permian they are going to face some cost pressures in the short-term, but we see that more cyclical.	0.991424	Return to shareholders
we bought back in as the shares looked attractive at the end of last year.	0.986496	Return to shareholders
that cash can then be used to buy businesses, generate growth and reward shareholders.	0.986423	Return to shareholders
it is just execution on the restructuring and not dropping the ball on the organic growth.	0.993555	Restructuring
the restructuring in the uk and other areas and so on, if that works, then the cash flow generation will ramp up materially.	0.991617	Restructuring
they are restructuring, they are focusing on the higher margin business, but that has cost a bit of top line and so they have got to return that to growth.	0.990271	Restructuring
they need to integrate all the businesses they bought and they need to demonstrate to the market that they did not overpay for all these assets - unattributed d it is just execution on the restructuring and not dropping the ball on the organic growth.	0.957231	Restructuring
i just want a rebase on the number.	0.938382	Restructuring
i cannot express these kind of views for compliance reasons.	0.842245	Regulation
in terms of profitability, they have been doing quite well and everything is fine, according to me.	0.990448	Profitability
they are very profitable there, so they could have increased competitive pressure in scandi.	0.987895	Profitability
we have improving free cash flow, primarily on the back of improving profitability on one hand, but also falling capex after they complete the plants, the old berendsen plants in the uk.	0.985849	Profitability
operating performance has been okay, but either they need to manage better expectations of the market and consensus, or they really need to start beating numbers again on the profit side, not just showing decent organic growth - melqart margin has been the big disappointment since the ipo.	0.983106	Profitability
i think the strategy is more to increase the profitability.	0.982398	Profitability
regarding the business, they are quite balanced between workwear, linen, etc.	0.982174	Product diversification
the market is less metro, so the externalisation of the activities, particularly in the workwear, could be a positive thing for the growth in brazil and in all latin america.	0.959193	Product diversification
regarding the business, they are quite balanced between workwear, linen, etc - midcap on balance investors and analysts regard elis operating performance positively and the management team is perceived to have a good track record for operational delivery.	0.952046	Product diversification
hygiene is regarded as an attractive opportunity by a number of respondents as it is less cyclical and capital intensive than the workwear and flat linen segments.	0.929292	Product diversification

the company also offer specific eco-labelled products, and in the uk they hold the carbon trust water standard.	0.914401	Product diversification
i think that the crown jewel of berendsen is even more valuable than france because it is similar amount of market share, but they have more pricing power and the organic growth has been stronger.	0.992525	Pricing power
it started quite quickly after they were ipod with the pricing pressure in france that was fairly strong.	0.991773	Pricing power
if they are unable to pass it through, there is a question around their pricing power and there is a question around the competitive dynamics, so that is another point.	0.990222	Pricing power
the attraction for us is, where you have got disciplined markets and you have got good pricing because you have only got two players or three players, then you tend to get quite good cash flow.	0.989641	Pricing power
after they were saying france is positive, we like france, it is doing well etc, two months afterwards they came out saying the pricing in france is difficult.	0.984540	Pricing power
the level of debt is also holding investors back from buying more stock now.	0.910799	Overhang
the long-lasting ownership from private equity put a bit too much constraint on what they could do in terms of consolidation, in terms of use of capital.	0.844337	Overhang
then i suppose the debt situation, which will hang around because they are not producing so much cash that they can pay it down that fast.	0.825415	Overhang
it is too high.	0.820515	Overhang
there is that angle.	0.816690	Overhang
perhaps it is more a question of communication in terms of m&a, to be less aggressive.	0.997684	Mergers and Acquisitions
it is becoming the leader in each country, at the right price of course - i am referring to m&a there.	0.997484	Mergers and Acquisitions
it is both the capex they have to deploy now, last year and this year, and maybe the worriedness around the deployment of capital in other m&a targets.	0.997231	Mergers and Acquisitions
these guys, they just want to keep on doing further m&a.	0.996468	Mergers and Acquisitions
as mentioned elsewhere in this report the current financial position is seen to preclude further m&a, which is viewed as a central plank of elis strategy.	0.995956	Mergers and Acquisitions
not on the management, but on the market, on this balance sheet structure, on the uk exposure, which is part of the business - lfde on the negatives, it is hard to know, berendsen might turn out to be brilliant or it might turn out to be a step too far.	0.992937	Market exposure
not on the management, but on the market, on this balance sheet structure, on the uk exposure, which is part of the business.	0.990325	Market exposure
now, due to the markets, it seems a bit high.	0.987591	Market exposure
secondly, maybe disclose more about capex and the working capital because that is an area of concern for the market.	0.986597	Market exposure

markets will penalise a business quite hard in a downturn and suddenly you become a very cyclical stock as opposed to what could be a very defensive business.	0.983093	Market exposure
we see margins trending up.	0.991612	Margin evolution
we have seen this evolution which was quite positive in terms of growth, in terms of margin and in terms of valuation for the stock.	0.978876	Margin evolution
then the question again, it is not their organic growth, but more can they deliver on their margin expansion and an improvement on the returns going forward or not?	0.974806	Margin evolution
we have seen this evolution which was quite positive in terms of growth, in terms of margin and in terms of valuation for the stock - unattributed b i think now they are much better diversified.	0.974368	Margin evolution
it is a good mix between the organic sales growth and the margin evolution.	0.973135	Margin evolution
experienced management team elis has an experienced management team, xavier martir in particular has worked in the business for many, many years.	0.996854	Management
it is a management team that is too focused on growth, not too focused on their own operations today, but has big things to do.	0.996032	Management
perception study - january 2019 management the management team is liked and are seen to be very good operators who manage their business well.	0.995597	Management
management is strong in terms of the way they think about things, at least.	0.995406	Management
management is strong in terms of the way they think about things, at least - permian leadership, i would say.	0.993074	Management
we will see how they are able to continue to grow in this different, maybe more complicated, macroeconomic context.	0.982857	Macro economics
they are good at m&a.	0.971235	Macro economics
there are cycles in the macro in every country.	0.937507	Macro economics
a lot of it is quite macro related in these individual markets and they have as good a view on that as anybody else really.	0.936576	Macro economics
there are macro factors that are near-term headwinds, plus the gilets jaunes in france.	0.932273	Macro economics
the leverage issue we have touched on.	0.993378	Leverage
leverage is one of these things where people think you should have more and more leverage and then they suddenly decide you should not have leverage and those two things can be a week apart when you look at the leverage profile in terms of time, it is a reasonable way out.	0.990141	Leverage
leverage is one of these things where people think you should have more and more leverage and then they suddenly decide you should not have leverage and those two things can be a week apart.	0.988416	Leverage
it is true that the leverage of elis is more than three times net debt to ebitda, but it is not an issuei would probably change my thesis if the company has an issue in terms of financing itself, but we know, and we had a discussion with the cfo, that they have no issue.	0.983995	Leverage

the overview is that the current share valuation, which is low on a pe and fcf yield basis (once capex normalises), is discounting all the negatives around the investment case of which the leverage is the principal issue and the investment market is unforgiving towards indebted companies currently.	0.983405	Leverage
as i said at the beginning, it is a leader.	0.997452	Leadership
we have been invested in elis because the company was run really properly, and we do not want them to change because of some investor who does not understand how the business has to be run - alken as i said at the beginning, it is a leader.	0.992589	Leadership
confident because of the whole industry and now that they are a leader.	0.989442	Leadership
now, in europe, the market is mature and they are the leader.	0.986538	Leadership
leadership, i would say.	0.981880	Leadership
also, showing to the market that they have got the balance sheet under control because interest rates are going up.	0.996545	Interest rates
when you have this question of the leverage in europe, this question of interest rates coming up, which has been pushed last year, you have several people shorting all the mid cap names with high leverage.	0.994147	Interest rates
of course, if interest rates increase, the market could be worried about that.	0.994023	Interest rates
that is actually a medium-term thing because we have had low interest rates now for a decade.	0.993617	Interest rates
the point is more market perception and with interest rates rising and maybe market slowdown, there are many fears around high indebtedness.	0.993383	Interest rates
working with ey the company have explored the environmental benefits of its rental and maintenance model with the potential to reduced water and non-renewable energy consumptions by nearly 50% compared to in- house solutions.	0.930411	Innovation
working with ey the company have explored the environmental benefits of its rental and maintenance model with the potential to reduced water and non-renewable energy consumptions by nearly 50% compared to in-house solutions.	0.927361	Innovation
it is a good idea that the board is something totally different.	0.924591	Innovation
we are almost entrepreneurial in that sense.	0.923248	Innovation
we have added recently.	0.852991	Innovation
this is how much is going into industrial plants, this is how much is going into other capex, this is how much of it is going into linen for the flat linen business.	0.995820	Industry specificities
bearing in mind that in spain, this is a more cyclical business because 70% is coming from flat linen and hospitality, so it is a different ball game - dynamo southern europe is more a worry because in spain and portugal they had very nice growth rate in the recent past, in the two past years maybe, so maybe the comparison basis is more difficult.	0.992373	Industry specificities

it is already a fairly consolidated market, it is a capital-intensive industry, so i think there is a lower risk of new market entrants and disruption on the competitive front.	0.990380	Industry specificities
this is an industry that is very good for m&a and to do the roll-outs, like they do - granular the strengths of the business are they are able to win market shares versus their competitors.	0.989712	Industry specificities
you have got a section of the business that is quite defensive because it is healthcare and whatever, industry, around the edges, but hospitality is a bit more cyclical.	0.986482	Industry specificities
it is an industry in consolidation, then consolidation brings higher margins, higher returns.	0.997068	Industry dynamic
meanwhile, they have a competitor in johnson, that is the service leader in the industry and it has been gaining market share.	0.992721	Industry dynamic
the latin american business is improving, but we know this is volatile, so it is good that they are penetrating in this market which is showing stronger growth.	0.992552	Industry dynamic
it is an industry-wide thing as well.	0.991403	Industry dynamic
rentokil is making a lot of acquisitions.	0.986870	Industry dynamic
i think it is going to change now that, for example, they have hired a new ir.	0.907388	Industry capacity
ideally, you would prefer to have an independent chairman, etc, but it is not that uncommon, particularly for companies of this size in france.	0.901816	Industry capacity
consolidate the market.	0.899567	Industry capacity
it is quite a mature industry.	0.884993	Industry capacity
now you have two people in the ir team, so it is a bit better and you do not need the cfo at every conference and every investor meeting.	0.849586	Industry capacity
we rank governance quite high at 7.5, so 7 or 8 is fine for me.	0.997584	Governance
we rank governance quite high at 7.5, so 7 or 8 is fine for me - lfde other than my earlier comment, i do not think i have anything further to add.	0.996608	Governance
governance is a very good mark.	0.991833	Governance
a very good rating, no issue with governance.	0.981209	Governance
many investors have no issues with corporate governance at elis, and a few say elis ranks highly on their governance metrics.	0.979508	Governance
as they grow outside europe, it should be higher growth, but it comes with a bit more geopolitical risk, currency risk and so on.	0.919647	Geopolitics
the prospects of a more challenging macro economic / geopolitical environment does not help.	0.901599	Geopolitics
then, on the other hand, consolidating on a country level, over time.	0.817285	Geopolitics
opening in a new country is probably not on the agenda.	0.762352	Geopolitics
margins, i do not think they have a huge leverage right now, but it depends on the country.	0.726813	Geopolitics
roce for scandinavia, roce for france, roce for spain, latin america, uk, germany and central europe.	0.994699	Geographical specificities

that is my main issue with them, in the uk - granular regarding the geography, the business is more diversified now after the different acquisitions they made this past three years.	0.989973	Geographical specificities
that is mainly because of the geographic mix with receivables being higher in places like brazil.	0.988060	Geographical specificities
then you have countries like spain, brazil, where operating margins are lower, so if they grow organically, in theory, that should drive through some pricing, volume growth, covering fixed costs.	0.987872	Geographical specificities
there is some argument to be made that the workwear business has higher margins and more value add, but it is more related to your scale in a particular geography.	0.987372	Geographical specificities
then we do believe there is a very strong, attractive story in their ability to consolidate the market and decrease their exposure to the french market and enlarge their exposure to new markets, new countries, which actually is exactly what they have been delivering until the acquisition of berendsen.	0.993737	Geographical diversification
14 0 1 2 3 4 5 6 7 culture accretive m&a geographic diversification high barriers to entry customer relationships market position management resilient model elis strengths - no.	0.993608	Geographical diversification
an increasingly diversified footprint the steady diversification of elis geographic footprint across europe has reduced reliance on the mature french market that had dominated the groups earnings pre-berendsen.	0.993252	Geographical diversification
there is a big story of market consolidation, which is the main driver in this investment case and get less and less exposure to the french market, which was their main market ten years ago - lfde post the ipo, the strategy was really to expand in other geographic markets.	0.992752	Geographical diversification
75 0 1 2 3 4 5 6 7 culture accretive m&a geographic diversification high barriers to entry customer relationships market position management resilient model elis strengths - no.	0.992653	Geographical diversification
these investors and analysts are cognisant of the markets aversion to high leverage at the moment reflecting concerns over the outlook for interest rates and the macro economic environment and this is reflected in the low rating for financial position of 5.3 out of 10.	0.993191	Financial Position
leverage driving the bear case the ranking of 5.3 out of 10 for financial position is the 5th lowest across all ip studies and compares to the average of 7.1.	0.991901	Financial Position
let us say six, because of the financial position.	0.990585	Financial Position
the spider graph below plots plots elis' financial position rating against other companies (a to bj) that comprise the ip index (red line).	0.990103	Financial Position
perception study - january 2019 financial position (cont.)	0.988063	Financial Position
not china - permian it is sensible.	0.965282	Environment
it is natural.	0.961222	Environment
uk & ireland, the main thing there is to get it back to positive growth.	0.941248	Environment

uk and ireland, low growth, so maybe around 1%.	0.935726	Environment
uk & ireland, obviously the margins are low, more lower, as catch-up.	0.919616	Environment
in terms of their strategy so far, the issue here is that they said we are going to consolidate markets and see margin expansion and i feel like that is being diluted a little bit.	0.993237	Dilution
then, on the other hand, consolidating on a country level, over time - unattributed c in terms of their strategy so far, the issue here is that they said we are going to consolidate markets and see margin expansion and i feel like that is being diluted a little bit.	0.986476	Dilution
then like i say, people are a bit worried there is this dilution to come from a big deal, which may well be absolutely the right thing to do, but people are i guess, a bit cautious to buy ahead of that - unattributed e sell-side mixed.	0.980051	Dilution
the fact that the growth profile, going outside france, is improving and diluting the french risk.	0.979252	Dilution
then like i say, people are a bit worried there is this dilution to come from a big deal, which may well be absolutely the right thing to do, but people are i guess, a bit cautious to buy ahead of that.	0.968249	Dilution
they made this huge deal with berendsen, leveraged the company more, and, on top of it, the market got materially worse.	0.975788	Deal structure
now, with the berendsen acquisition, it is well balanced.	0.974586	Deal structure
basically, they need to increase the hurdle rate to make acquisitions.	0.952245	Deal structure
keep doing small deals, rather than a very big structuring deal, leading to higher indebtedness and maybe capital increase.	0.936531	Deal structure
they can acquire quite expensive deals with dilution of returns on capital employed and that is absolutely fine with them.	0.936530	Deal structure
delivery, delivery, delivery.	0.960867	Cyclicality
they are going to face some cost pressures in the short-term, but we see that more cyclical.	0.952693	Cyclicality
europe is a dog of an end market, structurally, cyclically, in every way, shape or form brazil is a worse category - permian they are not that radically different.	0.931826	Cyclicality
europe is a dog of an end market, structurally, cyclically, in every way, shape or form.	0.926955	Cyclicality
it is a vicious circle.	0.905810	Cyclicality
obviously, sales will move with the fx.	0.940149	Currency
that is why i am saying now that i am a bit more worried about the cfo.	0.919850	Currency
perception study - january 2019 la financire de lechiquier (cont.)	0.911541	Currency
that would be my big picture view - moneta uk, that is a question mark.	0.909465	Currency
their roce is quite low overall.	0.834379	Currency

then there is the cost inflation, so they have to deal with some market topic.	0.998871	Cost inflation
the market is focused on financial leverage and now cost inflation.	0.997476	Cost inflation
near-term, cost inflation is an issue and the extent to which elis can pass this through.	0.995935	Cost inflation
they will have some constraints, in terms of cost, inflation of cost, so could be less reflecting in terms of margins.	0.995317	Cost inflation
next year, not really because there is some cost inflation, but going forward with berendsen back on track, margins should go up - bdl the ebitda margin is 31%, which is less relevant because they are pretty capex intensive, so you should focus on the ebit margin - dk partners we need to wait for what they are going to do because ultimately, what is going to happen.	0.989644	Cost inflation
there is competition, but the berendsen brand is a strong brand, so there is no reason why it should not continue - bdl scandinavia is probably the crown jewel.	0.996723	Competition
there is competition, but the berendsen brand is a strong brand, so there is no reason why it should not continue.	0.994466	Competition
the outcome, even in france where they are clearly the leader, depends on of course the competitive landscape and it can change.	0.993207	Competition
it was extraordinary really and it was really strange because it is not often you get companies talking about their competition.	0.990031	Competition
they claim it is competitive.	0.986603	Competition
the last negative is just in terms of raw materials; it does impact them, any movement in oil price or cotton price and stuff - franklin templeton we worry in general that elis is very focused on maximising its cost base.	0.993407	Commodity prices
we will probably pay attention to electricity prices and energy prices, oil, etc, because it will probably impact the group, but for now, no problem.	0.981195	Commodity prices
the last negative is just in terms of raw materials; it does impact them, any movement in oil price or cotton price and stuff.	0.980738	Commodity prices
these type of things, we can see it, but our first focus is in terms of cash flow.	0.998005	Cash-flow
it is a cash flow generator.	0.995718	Cash-flow
it is a cash flow question now.	0.994922	Cash-flow
their communication is clear in terms of what the targets are three years out, across the board, for the income statement, the balance sheet and the cash flow statement.	0.994186	Cash-flow
the cash flow, to prove that they can generate better cash flow - berenberg it is the situation in the uk.	0.992875	Cash-flow
he should try to understand what people are getting at and maybe there are things that he is doing that are not quite right - dynamo in terms of weaknesses, this is a high capex business, high capital-intensity business, a business that is prone to labour unrest.	0.997375	Capital intensity
in terms of weaknesses, this is a high capex business, high capital-intensity business, a business that is prone to labour unrest.	0.995510	Capital intensity
capital intensity may appear a little bit high, but there is also the impact of berendsen, which has underinvested recently, so we need to go to one year, i think, of high capex due to the berendsen investment, but it then should come back to more normal levels.	0.994593	Capital intensity

i suppose a weakness you could point to on the strategy is that it is capital-intensive, although that is a barrier, so it can have a reasonable level of debt.	0.994265	Capital intensity
as elis is really capital intensive, and it was exactly the contrary for berendsen, in a normalised environment the capex to sales ratio is really high.	0.993973	Capital intensity
organic growth and margins also important elis has a good business model and is undoubtedly a good operator in the supply of services to its customers and manages this well.	0.989217	Business model
it is a strong business model.	0.987696	Business model
it is very consistent with their cash flow and their business model.	0.983191	Business model
usually it should be quite a resilient business model, but when you have one-off events like that or, brexit in the uk, now that they have bought berendsen, that could shake a bit the resilience of the model.	0.979692	Business model
for us, it is not really a problem because we think it is the business model of the group and they generate enough free cash flow to pay back the debt and to decrease the debt ratio.	0.979385	Business model
high barriers to entry, high market share and the other thing is basically the sticky customer relationships - franklin templeton strengths.	0.985500	Barriers to entry
it creates barriers to entry because the company becomes bigger and bigger.	0.984485	Barriers to entry
you can see that with mainly distributors, like bunzl, but the barrier to entry is even higher, which is even more positive.	0.980245	Barriers to entry
it is a defensive company.	0.980168	Barriers to entry
first of all, the problem is that you do not have a long financial history with the company, not as a listed entity.	0.977125	Barriers to entry
in uk, i hope they could restore the growth base in the future.	0.830295	Backlog
i think it is fairly high when we look at how many years it would take them to repay that debt.	0.780751	Backlog
it is just that the uk, there is a lot to do.	0.692178	Backlog
this is not a growth company.	0.676738	Backlog
we also show the average of all the studies we have conducted (black line).	0.595866	Backlog